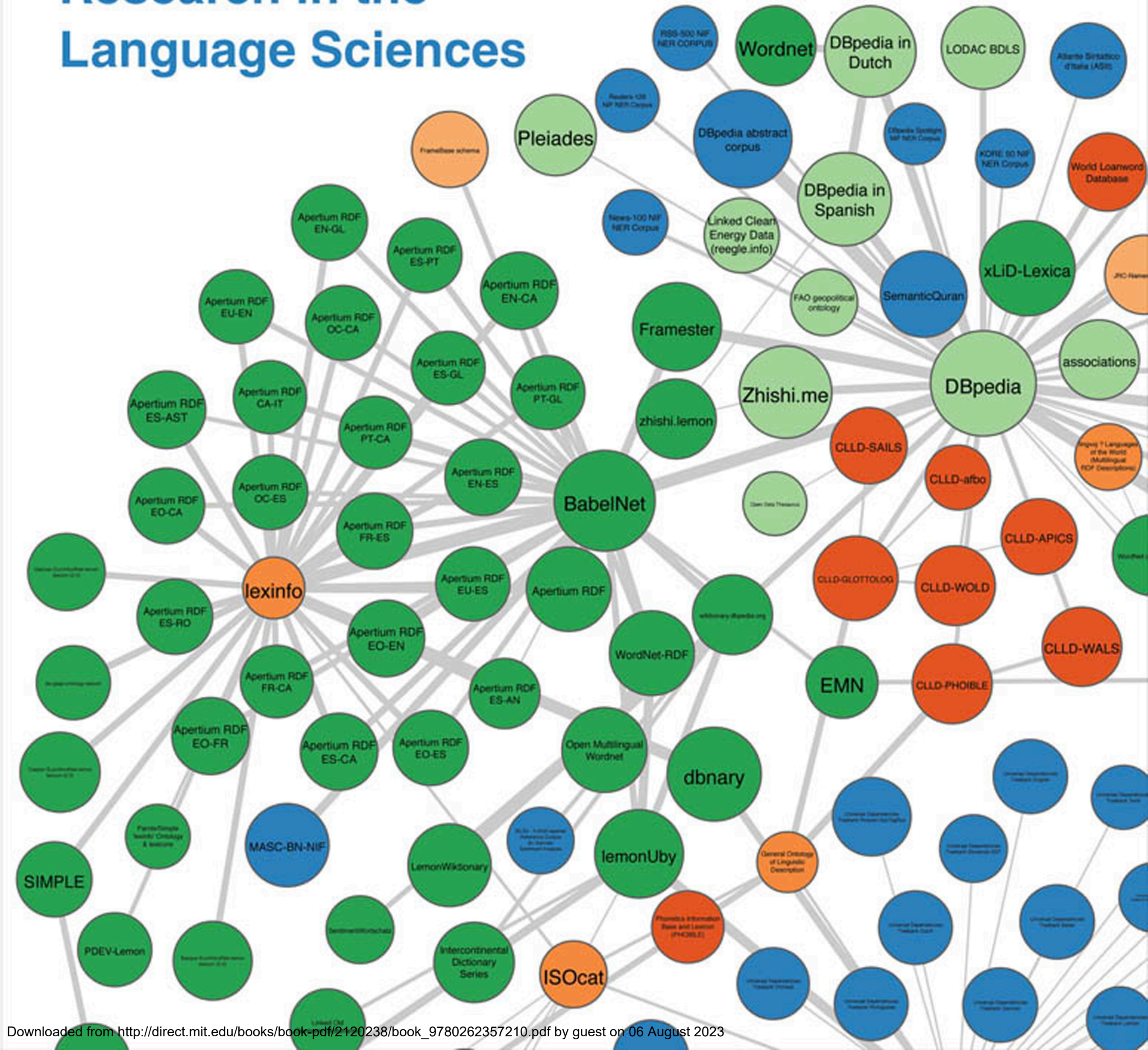
Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences

edited by

Antonio Pareja-Lora,
María Blume,
Barbara C. Lust,
and Christian Chiarcos



Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences

Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences

Edited by Antonio Pareja-Lora, María Blume, Barbara C. Lust, and Christian Chiarcos

The MIT Press
Cambridge, Massachusetts
London, England

© 2019 Massachusetts Institute of Technology

This work is subject to a Creative Commons CC BY-NC-ND license.

Subject to such license, all rights are reserved.

The Open Access edition of this book was published with generous support from the National Science Foundation (grant number BCS-1463196), Pontificia Universidad Católica del Perú, and Knowledge Unlatched.

This book was set in Times New Roman by Westchester Publishing Services. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Pareja-Lora, Antonio, editor. | Blume, María, editor. | Lust, Barbara C., 1941– editor. | Chiarcos, Christian, editor.
Title: Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences / edited by Antonio Pareja-Lora, María Blume, Barbara C. Lust, and Christian Chiarcos.
Description: Cambridge : MIT Press, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2019019588 | ISBN 9780262536257 (paperback)
Subjects: LCSH: Language and languages--Study and teaching. | Language and languages--Research. | Linked data.
Classification: LCC P53 .D398 2019 | DDC 025.06/4--dc23
LC record available at https://lccn.loc.gov/2019019588

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments vii

Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences: An Introduction ix
Barbara C. Lust, María Blume, Antonio Pareja-Lora, and Christian Chiarcos

1 **Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction 1**
Christian Chiarcos and Antonio Pareja-Lora

2 **Whither GOLD? 19**
D. Terence Langendoen

3 **Management, Sustainability, and Interoperability of Linguistic Annotations 25**
Nancy Ide

4 **Linguistic Linked Open Data and Under-Resourced Languages: From Collection to Application 39**
Steven Moran and Christian Chiarcos

5 **A Data Category Repository for Language Resources 69**
Kara Warburton and Sue Ellen Wright

6 **Describing Research Data with CMDI—Challenges to Establish Contact with Linked Open Data 99**
Thorsten Trippel and Claus Zinn

7 **Expressing Language Resource Metadata as Linked Data: The Case of the Open Language Archives Community 117**
Gary F. Simons and Steven Bird

8 TalkBank Resources for Psycholinguistic Analysis and Clinical Practice 131
Nan Bernstein Ratner and Brian MacWhinney

9 Enabling New Collaboration and Research Capabilities in Language Sciences: Management of Language Acquisition Data and Metadata with the Data Transcription and Analysis Tool 151
María Blume, Antonio Pareja-Lora, Suzanne Flynn, Claire Foley, Ted Caldwell, James Reidy, Jonathan Masci, and Barbara Lust

10 Challenges for the Development of Linked Open Data for Research in Multilingualism 185
María Blume, Isabelle Barrière, Cristina Dye, and Carissa Kang

11 Research Libraries as Partners in Ensuring the Sustainability of E-science Collaborations 201
Oya Y. Rieger

List of Contributors 213
Author Index 221
Thematic Index 225

Acknowledgments

This volume and the 2015 workshop, Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences, convened at the University of Chicago in July 2015, were funded by the National Science Foundation with a grant (Award ID 1463196) given to Barbara C. Lust, María Blume, and Antonio Pareja-Lora. Publication was further supported by a grant from the Cornell University Library and supplemented by the Cornell Institute for Social Science and the Cornell Cognitive Science Program, as well as the Department of Humanities at the Pontificia Universidad Católica del Perú and Knowledge Unlatched.

We thank William J. Badecker, NSF Linguistics Program director, whose advice has been invaluable during all stages of this project, and Marc Lowenthal and Anthony Zannino at MIT Press, who guided us to publication. We thank Amy Brand, director of MIT Press, whose pursuit of the development of scholarly communication in the digital age provided support for our project. The Cornell University Library, through Oya Rieger, provided continual advice and support, as well as a critical dimension of library-researcher relations, continuing the early vision of previous Mann Library director, Janet McCue. Emily Bernardski provided key support and coordination for the workshop, as did Carissa Kang and Jonathan Masci, our student support team. Our editors Michelle Melanson and Rebecca Rich Goldweber provided invaluable assistance in volume publication. James Gair provided continual support throughout.

Previous supporters for the development of the LLOD vision and related research that established the foundations for this project are María Blume and Barbara Lust (2008), Transforming the Primary Research Process Through Cybertool Dissemination: an Implementation of a Virtual Center for the Study of Language Acquisition. NSF OCI-0753415; Janet McCue and Barbara Lust (2004), National Science Foundation Award: Planning Information Infrastructure Through a New Library-Research Partnership (SGER = Small Grant for Exploratory Research NSF 0437603); and Barbara Lust (2003) Planning Grant: a Virtual Center for Child Language Acquisition Research, National Science Foundation, NSF BCS-0126546. Additional support has been provided by the American Institute for Sri Lankan Studies, the Cornell University Einaudi Center, Cornell

University Faculty Innovation in Teaching Awards, the Cornell Institute for Social and Economic Research (CISER), and the Cornell Institute for Social Sciences.

Finally, we gratefully acknowledge the welcome and continual collaboration of other founding members of the Virtual Center for the Study of Language Acquisition (VCLA) for supporting the vision represented in this volume: Suzanne Flynn (MIT, USA); Qi Wang, Marianella Casasola, and Claire Cardie (Cornell University, USA); Elise Temple (The Nielsen Company, USA); Liliana Sánchez (Rutgers University at New Brunswick, USA); Jennifer Austin (Rutgers University at Newark, USA); YuChin Chien (California State University at San Bernardino, USA); and Usha Lakshmanan (Southern Illinois University at Carbondale, USA). We greatly appreciate the collaboration of scholars who are VCLA affiliates, including Sujin Yang (Ewha Womans University, South Korea); Gita Martohardjono (City University of New York Graduate Center and Queens College, USA); Valerie Shafer (City University of New York, USA); Isabelle Barrière (Long Island University–Brooklyn, USA); Cristina Dye (Newcastle University, UK); Yarden Kedar (Beit Berl College, Israel); Joy Hirsch (Columbia University, USA); Sarah Callahan (Assessment Technology Inc., USA); Kwee Ock Lee (Kyungsung University, South Korea); R. Amritavalli (Central Institute of English and Foreign Languages, India); and A. Usha Rani (Osmania University, India).

Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences: An Introduction

Barbara C. Lust, María Blume, Antonio Pareja-Lora, and Christian Chiarcos

This volume arose out of a workshop, *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, held under the auspices of the Linguistic Society of America (LSA) Summer Institute at the University of Chicago in July 2015. The workshop was organized by Barbara Lust, Antonio Pareja-Lora, and María Blume, with the support of the National Science Foundation (NSF 1463196), supplemented by support from Cornell University's Institute for Social Sciences and Cognitive Science program. The collection of papers in this volume results from that workshop. Publication was further supported by the Cornell University Library, the Department of Humanities at the Pontificia Universidad Católica del Perú and Knowledge Unlatched.

The workshop was energized by the transformation in science scholarship that has developed over recent decades and that was envisioned by the National Science Foundation's Blue-Ribbon Advisory Panel on CyberInfrastructure (Atkins et al. 2003, reviewed and assessed by Borgman 2007). Empowered by the internet, the current digital age opened unprecedented opportunities for storing, disseminating, sharing, and manipulating large and complex amounts of data to become *open* and *linked* (Berners-Lee 2009; Chiarcos, Hellmann, and Nordhoff 2012). The more that each data singleton can be significantly *interlinked*, the more powerful and useful it becomes, enabling scholars to pursue new and advanced questions. The more that data are linked, and the more that datasets and data providers are included in this linking, the more that researchers within and across disciplines can partner, based on shared data, thereby empowering more powerful research questions. Today, many of the sciences, including the social sciences, are being transformed by these developments. Yet, converting disparate and self-contained databases (data silos) into interlinked resources to facilitate co-operation and synergies between academic researchers is only one aspect of that transformation. It is accompanied by corresponding developments in politics and society: President Barack Obama's Executive Order on Open Data[1] as well as the current federal funding agency requirements for data management and data sharing plans assumed this transformation. The concept of Open Data, which is achieved via exploiting the internet and cloud resources, offers great promise across the sciences, and in fact, is recently becoming required by federal mandates in conjunction with research funding. This science-wide energy cohered with an active concern discussed at the LSA 2015 Summer Institute, reflected in its theme, “Linguistic Theory in a World of Big Data,” highlighting “a growing interest within the field of linguistics to test theory with increasingly larger data sets…”[2] that integrate across both large and diverse language data sources.

The LSA workshop was also specifically energized by the Open Linguistics Working Group (OWLG) of Open Knowledge International.[3] Open Knowledge International represents a worldwide network of people aiming to demonstrate the value of Open Data to society. As a nonprofit organization, it brings together enthusiasts, providers, and consumers of Open Data to facilitate advocacy, technology, and training with the goal of unlocking information and empowering people to create and share knowledge on this basis. A number of working groups take a more specific focus, and since its foundation in 2010 the working group on Open Linguistics has strived to bring together researchers, students, and practitioners from various branches of the language sciences (from academia, applied linguistics, lexicography) and computer science (natural language processing, knowledge representation, artificial intelligence/Semantic Web, localization industry) in a shared concern for developing, promoting, and using open language resources (Chiarcos, Hellmann, and Nordhoff 2012; Chiarcos et al. 2013).

The principal activities of the Open Linguistics Working Group include the organization of workshops, most notably the Linked Data in Linguistics workshop series that aims to discuss types of resources; strategies to address issues of interoperability between them; protocols to distribute, access, and integrate this information; and technologies and infrastructures developed on this basis. In this context, Chiarcos et al. (2013: i) observed that “[t]he lack of interoperability between linguistic and language resources represents a major challenge that needs to be addressed if information from different sources is to be combined,” but that “commonly accepted strategies to distribute, access and integrate their information have yet to be established, and technologies and infrastructures to address both aspects are still under development.”

In response to this challenge, the Open Linguistics Working Group engaged in a joint effort to adopt the Linked Open Data (LOD) paradigm as a technical means to facilitate the use, reuse, harmonization, and interoperability of language resources. In 2011, a Linguistic Linked Open Data (LLOD) cloud was envisioned, drafted, and described, and as a result of a datathon held on Multilingual Linked Open Data for Enterprises (MLODE-2012, in Leipzig, Germany), a core dataset and the first LLOD diagram were presented and have continued to grow.[4] Since August 2014, these activities have been acknowledged by giving *linguistics* the status of a top-level category in the LOD cloud diagram.[5]

Since its foundation, OWLG activities and LLOD development have been supported by various international research projects,[6] many of which were funded by the European Union as a means to reduce language and knowledge barriers in Europe’s digital Single Market.

While heterogeneity in languages and language resources is a perpetual global challenge, this support led to a natural focus of LLOD activities throughout Europe. At present, however, the LLOD initiative has still not widely reached relevant scholars in the language sciences in the United States and the Americas. In particular, subareas of the language sciences, which are providers of both research knowledge or content—for example, the subarea of psycholinguistic research such as that involved in the study of language acquisition—have neither integrated the LLOD agenda nor been deeply integrated into it. Where advances have in fact been made in various areas related to developments in interoperability, they have often failed to become widely known across the fields of the language sciences.

The purpose of this volume, as of the workshop, is to advance an international infrastructure for scholarship in the field of the language sciences—one that will help to expand LOD power for the language sciences. The chapters within do so by merging active research demands in the field of language acquisition with technical advances occurring now in the development of data interoperability. Specifically, the authors' purpose was to cultivate a multidisciplinary international community that could collaboratively address the promises of a Linked Open Data dimension in both linguistics and the language sciences, and then could begin to exploit this community in order to meet the conceptual and technical challenges necessary for the attainment of LLOD. This volume, like the workshop that inspired it, convenes research and analysis of a multidisciplinary group of international scholars who are currently engaged in meeting both technical and conceptual challenges involved in linking data for collaborative research. By this convergence, the authors hope to develop communication and advanced synergy between scholars with active research needs and those developing technical capacity to enable solutions through the various multifaceted demands of LOD, including both operative engineering advances and ontologies enabling interoperable language data. These two domains of scholarship, in fact, rarely overlap—a gap that must be bridged if a LOD agenda in the language sciences can ever be advanced.

In pursuing this purpose, we recognize that "data scholarship is rife with tensions between the social and the technical [..., yet] rarely can these factors be separated [... as] the social and the technical aspects of scholarship are inseparable ..." (Borgman 2015, 35). Development of technical resources is critically necessary to the LOD vision; technology of cyberinfrastructure "makes data creation possible" (Borgman 2015, 35). At the same time, "the ability to imagine what data might be gathered" (Borgman 2015, 35)—and why—must both ground and empower this technology. It must give these data meaning and purpose.

In this volume, we seek to address the tension between social and technical aspects of the LOD vision by bringing scholars in the technologies of information science—necessary to LOD—together with scholars in the language sciences who are confronted with real research needs for data representation, dissemination, and related collaboration. The social–technological integration we pursue allows researchers of both areas to become aware of the other's developments and challenges. Specifically, it enables them to share, first, cutting-edge developments in the technology of Linked Open Data, and second, cutting-edge research in the language sciences that are generating content, where researchers are seeking to advance a LOD agenda. This integration is necessary to support collaborative, cross-linguistic research involving calibrated data sharing enabled by current and developing technology of a networked environment. The community present at the workshop and represented in this book includes:

- Researchers in the content area/multilingualism in children, since this is the area that we chose as our concrete test case example, where data merging and sharing are motivated by ongoing research questions
- Researchers and developers working directly in the development of the LOD cloud and, whenever possible, interested also in generating and enhancing the Linguistic Linked Open Data (LLOD) cloud
- Computer engineers and researchers with some experience in solving interoperability problems (e.g., ontological engineers)
- Linguists and computational scientists with some expertise in developing computational models of language and/or linguistics and in language/linguistic annotations (e.g., computational linguists)
- Experts in the area of standards and/or standardization, of both the language-related and the knowledge representation domains. On the one hand, language-related standard experts came mainly from the ISO/TC 37 technical committee, which deals with terminology, knowledge management, and language resources. On the other hand, knowledge representation standardization experts need to be aware of the different recommendations of the World Wide Web Consortium (W3C), such as XML, RDF, OWL, which have been essential in the development of the Semantic Web and/or the LOD cloud.

We chose the area of acquisition of language, including multilingualism, as our focus, because it requires highly complex and diverse language datasets, cross-linguistic analyses, and urgent collaborative research. We hope that this volume's examples of scholarly convergence of technical and social science in the area of LOD can provide models for advances in other areas of science, as well.

Challenges

The Concepts

The very concepts underlying the LLOD vision are extremely complex. Uncertainties and disagreements persist, regarding what constitutes *data* and what constitutes *open* (Borgman 2015; Chiarcos and Pareja-Lora, this volume).

Moreover, language data itself provides particular complexities. Natural language is multidimensional and involves several levels of representation—phonological, syntactic, and semantic, as well as pragmatic, for example—and in turn each of these dimensions is analyzed in the scientific study of its components. Language data arise aurally or (in the case of sign languages) visually, and are represented in multiple dimensions, varying from acoustic to written. Cross-language variation exponentially increases the complexity of representation and analysis. Data can be derived either experimentally or observationally. The largest driving questions in the language sciences, such as *what is biologically programmed?* versus *what is experientially derived?*, involve cross-disciplinary investigations that include neuroscience.

Technology and Systems

Technological advances require interoperability of systems and services, and standards relate to such interoperability (Borgman 2007). At the same time, for the creation of Linked Open Data networks "expansion depends more on the consistency of data description, access arrangements, and intellectual property agreements than on technological advances" (Borgman 2007, 10).

The Challenges Now in the Language Sciences

The realistic accomplishment of a vision of LOD requires confronting several challenges, and these require a diverse yet integrated community to address them.

1. A priori, language sciences databases and local infrastructures are not interoperable, owing, to the various conceptualizations and/or database schemas used to acquire, store, and manage their data, and to the actual ways in which they are stored and managed (in a database and/or different types of files with different formats).
2. Various linguistic theories can be applied for data description and analysis. This creates a need to interface theoretical vocabularies (e.g., by means of ontologies and ontology mappings) when merging and linking different language databases and/or resources.
3. Annotation schemas resulting from specific ontologies can vary widely, with specific research agendas requiring precise and specific theory-driven data markup and with general knowledge provider frameworks that must interface with these theoretical vocabularies in a computationally practical manner (Pareja-Lora, Blume, and Lust 2013 begin to approach this challenge).
4. Data proprietors are reluctant to share a very valuable resource that, in most cases, has been developed after devoting considerable human, economic, and time resources, without established principles of collaboration (Ledford 2008).
5. Legal issues arise, for example, if private and often confidential human-subject data either are shared with institutions other than the one where initial protection of human subjects was approved by local institutional review boards, or are made openly accessible, such that profit-driven entities can make use of data gathered solely for a nonprofit purpose.

6. Cybertools are necessary to provide individual researchers with infrastructure for creating data in a form that can ultimately become efficiently interoperable.
7. In interdisciplinary contexts, scholars from varying areas of research and development often use different terms to refer to the same concepts and ideas, which challenges interdisciplinary work in general and technology's integration with specific research domains in particular. The very nature of data can vary across disciplines (e.g., neural data arising from neuroscience).
8. Bird and Simons (2003) note an important goal for resources that seek to make language data widely available: to maximize the benefit of the resource to as wide a community as possible while at the same time protecting what they term "sensitivities." Maximizing the benefits would include adding useful technological features (e.g., improving export functionality), both across data platforms and from a project housed at a particular member's lab to the open linked network and back.
9. Researchers need to retain some level of ownership of their data (e.g., participating in various uses of the data).
10. Research participants are entitled to confidentiality and thus to the protection of identifying information (cf. Blume and Lust 2017). Some kinds of leveled protections will be necessary for wide dissemination. The DOBES Programme[7] has already addressed some of these issues, as has the Cornell Institute for Social and Economic Research in working with census data.
11. The LLOD vision of shared research infrastructure and data confronts challenges of sustainability, as it does in other sciences (Berman and Cerf 2013). Who will maintain long-term servers, and how will that investment be supported?

In sum, the field of linguistics and the language sciences is challenged now to develop "(1) an infrastructure of collaboration (Ledford 2008); (2) standardized tools and best practices which can be shared while at the same time allowing unique methods by individual researchers; (3) infrastructure for data storage, management, dissemination and access, including means for interfacing databases that differ in both type and format; (4) preservation and 'portability' of data and related materials (Bird and Simons 2003, NSF 2007); and (5) changes to the ways in which we educate our students and train new researchers in scientific methods" (Blume and Lust 2012, 1).

Advances to Date

The language sciences have made advances on several dimensions required for confronting the field's data-intensive demands. Scholars represented in this volume report the current state of the art along these dimensions.

For example, with regard to the infrastructure of knowledge dissemination, the Open Language Archives Community (OLAC)[8] has made advances in facilitating access to language archives worldwide. Metadata development has progressed through language documentation initiatives, which attempt to confront not only data collection but also data interfacing with data and analyses across languages (e.g., Grenoble and Furbee 2010; Good 2002). As an initiative of E-MELD (Electronic Metastructure for Endangered Languages),[9] the GOLD ontology (General Ontology for Linguistic Description)[10] has advanced in development of a shared vocabulary for interfacing linguistic descriptions (mainly morphosyntactic and syntactic annotations) for language documentation. This endeavor has also been dealt with and extended to other linguistic levels in other ontological frameworks and/or models for linguistic annotation, such as OntoTag and OntoLingAnnot (Aguado de Cea et al. 2004; Pareja-Lora and Aguado de Cea 2010; Pareja-Lora 2012a, 2012b, 2012c, 2013, 2014, 2016a, 2016b; Pareja-Lora, Blume, and Lust 2013). Besides, this effort has been met with similar undertakings for natural language processing and lexicography (the ISO/TC37 Data Category Registry ISOcat, and its successor DatCatInfo)[11] and corpus linguistics (the Ontologies of Linguistic Annotation, OLiA).[12] The challenge of language data portability has been articulated by Bird and Simons (2003). On local levels, several initiatives of data sharing of various forms have developed, for example the Penn Treebank, serving various NLP projects in computer science (e.g., Marcus, Santorini, and Marcinkiewicz 1994) and computational linguistics and various corpus linguistics endeavors that require large amounts of data in the search for distributions and frequencies of language phenomena (e.g., Biber, Conrad, and Reppen 1998), as well as several endangered language initiatives (E-MELD, DOBES).[13] In the child language field, CHILDES[14] (MacWhinney and Snow 1985) has developed mechanisms for distribution of language data. The LinguistList[15] and the CHILDES mailing list both cultivate knowledge exchange. Psychologists are actively confronting issues of database use (Johnson 2001; Johnson and Sabourin 2001) and data replicability (e.g., Johnson 2014). The VCLA (Virtual Center for the Study of Language Acquisition) has developed the Data Transcription and Analysis (DTA) tool (cf. this volume).

Contributions in This Volume

The focus of this collaborative volume is to demonstrate and to illustrate the potential of Linked Open Data technologies in our area of research. While, naturally and most obviously, the infrastructural solutions developed on this basis facilitate exchange and reuse of Open Data, openness is not a requirement for the technology per se—nor is it for the collections assembled herein. Yet, with more and more concrete applications constantly being developed, we expect an increased readiness to commit to publish Linked Data as Open Data—and to publish Open Data as Linked Data.

Chiarcos and Pareja-Lora in chapter 1 introduce the reader to the basic concepts underlying the projects reported in this book, Open Data, Linked Data, Linked Open Data, and Linked Open Data in Linguistics, as well as the historical and recent development of LLOD.

Ide (chapter 3) explains how managing and processing the resources created by Linguistic Linked Open Data (LLOD) requires standardized, accessible applications that are interoperable with a vast array of other platforms and services. Such interoperability must be achieved, she argues, on several levels: physical formats must be compatible to effect syntactic interoperability, while models of linguistic objects and features must be harmonized to achieve semantic interoperability among applications and the data they process. She shows how INTEROP/SILT and The Language Application Grid have addressed the interoperability problem and have both proposed and then implemented some solutions, and also have identified best practices and recommendations for representing and exchanging linguistic data in linked or other form.

Chapters 2, 4, 5, 6, and 7 present many of the efforts of data description and annotation through the creation of ontologies, metadata, and repositories that seek to make Open Data sharable, comparable, and reusable. Langendoen (chapter 2) presents a history of the General Ontology for Linguistic Description (GOLD) and analyzes challenges involved in its development into a Digital Infrastructure that supports Linguistic Inquiry (DILI). The GOLD project aims to provide access to large amounts of digital data in various formats about many languages, data that are relevant to many different areas of research and application both within and outside of linguistics; to facilitate the comparison, combination, and analysis of data across media, languages, and subdisciplines; and also to support seamless collaboration across space, time, (sub)disciplines, and theoretical perspectives. Moran and Chiarcos (chapter 4) describe the application of LLOD technology to the publication and dissemination of linguistic data from under-resourced language data. They argue for the importance of Linked Data for encoding, sharing, and disseminating such data, while they discuss aspects of resource integration. Warburton and Wright (chapter 5) present DatCatInfo, an online resource for data categories used to document digital language resources, which replaces the earlier ISOcat.org Data Category Registry with a Data Category Repository developed in the terminology management system named TermWeb. This chapter both exemplifies and details the challenges and procedures for migrating original data category specifications to a new environment. Trippel and Zinn (chapter 6) describe the CLARIN research infrastructure, which offers researchers access to a wide range of language-related research data and tools. It aims to develop a common metadata framework that makes it possible to describe all types of resources to a fine-grained level of detail, paying tribute to both their specific characteristics and the requirements of the many social sciences communities. They stress the need for an initial data curation step and describe the connection of CMDI-based metadata to existing Linked Data, then consider how these data can be converted to bibliographic metadata standards and entered into library catalogs. They describe first steps to convert CMDI-based metadata to RDF. They also discuss how the initial grassroots approach of CMDI makes it difficult to fully link its metadata to other Semantic Web datasets, and they consider steps toward extending their Component MetaData Infrastructure's (CMDI) semantic interoperability beyond the social sciences and humanities. Simons and Bird (chapter 7) describe the Open Language Archives Community (OLAC), an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources that aggregates a union catalog of all the resources held by the participating institutions. OLAC has developed standards for expressing and exchanging the metadata records that describe the holdings of an archive. The authors explore the application of Linked Data to the problem of describing language resources in the context of OLAC.

Chapters 8, 9, and 10 are devoted to the actual challenges of representing research data in the field of language acquisition and of developing online resources in support of Linked Open Data. Bernstein Ratner and MacWhinney (chapter 8) describe how the TalkBank system has complied with certain requirements for Linguistic Linked Open Data and is currently developing methods for linkage and comparisons between corpora based on automatically computed quantitative measures. They provide examples of such measures using the KIDEVAL program. Blume et al. (chapter 9) describe development of the Data Transcription and Analysis tool (DTA tool), a web-based infrastructure that promotes strong metadata and data management and enables collaborative research with language data, while allowing for fine-grained, cross-linguistic comparison of linguistic phenomena. They explore technical challenges of this tool and exemplify their attempt to lay the foundations for development of a future, broad, Linked Open Data framework for collaborative research in the language sciences. Blume et al. (chapter 10) address the complex requirements for conducting research with multilingual populations (which characterize more than half the world's population) and sketch the challenges for the development of Linguistic Linked Open Data (LLOD) in this field. Research in this area of multilingualism, like all language research, requires the collection of metadata that are detailed and transparent enough to allow for replication and calibrated collaboration, but such research poses extra challenges for data markup.

Finally, Rieger (chapter 11) proposes that creating an open and linked linguistics research data infrastructure must entail a seamless network of content, technologies, policies, expertise, and practices, and doing so requires that such a scholarly organization be viewed as an enterprise that needs to be not only built but maintained, improved, assessed, and promoted over time. She discusses the potential role of research libraries as partners in fostering open science, based on Cornell University's experience in running arXiv, a model scientific preprints repository. It was important to the editors of this book that it had an open access version in digital format. A printed version is also available for purchase. However, be advised that figures in color are only available in the open access version. Readers can find it here: https://mitpress.mit.edu/.

Challenges for the Future

In general, although the field of the language sciences has made some advances on several dimensions required for confronting the field's data-intensive demands, as we have reviewed above and while the papers in this volume demonstrate many of those advances, the transformation of traditional, as well as current, resources into a Linked (Open) Data resource remains a far-from-straightforward process, as several papers acknowledge; indeed, to a large degree that transformation remains a vision for the future. The development of the LOD cloud remains in its infancy, and many details require extensive in-depth discussion before solutions can be implemented. It is hoped that the cross-field discussion initiated by the multidimensional research scholars presenting in this volume will help to cultivate and nurture what is necessary now, even as we pursue this vision.

Notes

1. https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.
2. https://lsa2015.uchicago.edu/.
3. https://okfn.org/.
4. http://linguistic-lod.org/.
5. https://lod-cloud.net.
6. Selected European research and innovation actions include LOD2 (11 EU countries + Korea, 2010–2014), MONNET (5 EU countries, 2010–2013), LIDER (5 EU countries, 2013–2015), QTLeap (6 EU countries, 2013–2016), FREME (6 EU countries, 2015–2017), Prêt-à-LLOD (6 EU countries, 2019–2021). Independently from these technology-centered projects, a number of larger scale projects from the humanities are based on LLOD technology, for example the ERC-funded research groups POSTDATA (on European poetry, 2015–2020) and Linking Latin (on Latin philology, 2018–2023), the research group Linked Open Dictionaries (on language contact studies, 2015–2020, funded by the German Federal Ministry of Education and Science), or the Trans-Atlantic Platform project MTAAC (on cuneiform studies, 2017–2019, funded by NEH, the Canadian SSHRC, and the German DFG).
7. http://dobes.mpi.nl/access_registration/.
8. http://www.language-archives.org/.
9. http://www.emeld.org/index.cfm.
10. http://linguistics-ontology.org/.
11. http://www.isocat.org/, resp. http://www.datcatinfo.net.
12. http://purl.org/olia/.
13. http://dobes.mpi.nl/access_registration/.
14. https://childes.talkbank.org/.
15. https://linguistlist.org/indexfd.cfm.

References

Aguado de Cea, G., I. Álvarez de Mon, A. Gómez-Pérez, and A. Pareja-Lora. 2004. "OntoTag's Linguistic Ontologies: Improving Semantic Web Annotations for a Better Language Understanding in Machines." In *Proceedings of the International Conference on Information Technology: Coding and Computing, 2004 (ITCC 2004)*, Vol. 2, 124–128. Las Vegas, Nevada, USA.

Atkins, D. E., K. K. Droegemeier, S. I. Feldman, H. García-Molina, M. L. Klein, D. G. Messerschimdtt, P. Messina, J. P. Ostriker, and M. H. Wright. 2003. "Revolutionizing Science and Engineering: Report of the National Science Foundation Blue-Ribbon Advisory Panel on CyberInfrastructure." January 2003. http://www.nsf.gov/cise/sci/reports/atkins.pdf.

Berman, F., and V. Cerf. 2013. "Who Will Pay for Public Access to Research Data?" *Science* 341:616–617.

Berners-Lee, T. March 2009. TED Talk video; Berners-Lee on the next web. http://www.ted.com/talks/lang/en/tim_berners_lee_on_the_next_web.html.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bird, S., and G. Simons. 2003. "Seven Dimensions of Portability for Language Documentation and Description." *Language* 79 (3): 557–582.

Blume, M., and B. C. Lust. 2012. "First Steps in Transforming the Primary Research Process through a Virtual Linguistic Lab for the Study of Language Acquisition and Use: Challenges and Accomplishments." *Journal of Computational Science Education (JOCSE*) 3 (1): 34–46.

Blume, M., and B. C. Lust. 2017. *Research Methods in Language Acquisition. Principles, Procedures and Practices*. Mouton de Gruyter and American Psychological Association.

Borgman, C. L. 2007. *Scholarship in the Digital Age*. Cambridge, MA: MIT Press.

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Chiarcos, C., P. Cimiano, T. Declerck, and J. McCrae, eds. 2013. *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and other Language Data*. Pisa: Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-5500.

Chiarcos, C., S. Hellmann, and S. Nordhoff. 2012. "The Open Linguistics Working Group of the Open Knowledge Foundation." In *Linked Data in Linguistics: Representing Language Data and Language Metadata*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 7–14. Berlin/Heidelberg: Springer.

CHILDES. https://childes.talkbank.org/.

DOBES Programme http://dobes.mpi.nl/access_registration/.

Electronic Metastructure for Endangered Languages (E-MELD) http://www.emeld.org/index.cfm.

Exec. Order No. 13642. 3 C.F.R. p. 244–246. 2013. Making open and machine readable the new default for government information. https://www.gpo.gov/fdsys/pkg/CFR-2014…/CFR-2014-title3-vol1-eo13642.pdf.

General Ontology for Linguistic Description (GOLD). http://linguistics-ontology.org/.

Good, J. 2002. A gentle introduction to metadata. http://www.language-archives.org/documents/gentle-intro.html.

Grenoble, L. A., and N. L. Furbee, eds. 2010. *Language Documentation: Practice and Values*. Amsterdam: John Benjamins.

Johnson, D. 2001. "Three Ways to Use Databases as Tools for Psychological Research." *APS Observer* 14 (10): 7–8.

Johnson, D., and M. Sabourin. 2001. "Universally Accessible Databases in the Advancement of Knowledge from Psychological Research." *International Journal of Psychology* 36(3):212–220.

Johnson, G. 2014. "New Truths That Only One Can See." *New York Times* January 20.

Linked Open Data in Linguistics (LLOD). http://linguistic-lod.org/.

Ledford, H. 2008. "With All Good Intentions." *Nature* 452 (10): 682–684.

"Let Data Speak to Data. 2005. *Nature* 438 (7068): 531.

LinguistList, The. https://linguistlist.org/indexfd.cfm.

MacWhinney, B., and C. Snow. 1985. "The Child Language Data Exchange System." *Journal of Child Language* 12:271–296.

Marcus, M., B. Santorini, and M. Marcinkiewicz. 1994. "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics* 19 (2): 313–330.

National Science Foundation. 2007. "NSF's Cyberinfrastructure Vision for 21st century Discovery." September 26. http://www.nsf.gov/pubs/2007/nsf0728/index.jsp.

Open Knowledge Foundation (OKFN). https://okfn.org/.

Open Language Archives Community (OLAC). http://www.language-archives.org/.

Pareja-Lora, A. 2012a. *Providing Linked Linguistic and Semantic Web Annotations: The OntoTag Hybrid Annotation Model*. Saarbrücken: LAP–LAMBERT Academic Publishing.

Pareja-Lora, A. 2012b. "OntoLingAnnot's Ontologies: Facilitating Interoperable Linguistic Annotation up to the Pragmatic Level." In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, 117–127. Berlin, Heidelberg: Springer Verlag. DOI: 10.1007/978-3-642-28249-2_12; http://link.springer.com/chapter/10.1007/978-3-642-28249-2_12.

Pareja-Lora, A. 2012c. "OntoLingAnnot's LRO: An Ontology of Linguistic Relations." In *Proceedings of the 10th Terminology and Knowledge Engineering Conference—New Frontiers in the Constructive Symbiosis of Terminology and Knowledge Engineering*, 49–64. Madrid: Universidad Politécnica de Madrid.

Pareja-Lora, A. 2013. "An Ontology-Driven Methodology to Reuse, Link and Merge Terminological and Language Resources." In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence—Terminology for a Networked Society: Recent Advances in Multilingual Knowledge-based Resources*, 189–196. Paris (Villetaneuse), France. https://lipn.univ-paris13.fr/tia2013/Proceedings/actesTIA2013.pdf#page=191.

Pareja-Lora, A. 2014. "The Pragmatic Level of OntoLingAnnot's Ontologies and Their Use in Pragmatic Annotation for Language Teaching." In *Languages for Specific Purposes in the Digital Era. Series: Educational Linguistics*, Vol. 19, 323–344. Switzerland: Springer International Publishing. DOI: 10.1007/978-3-319-02222-2_15; http://link.springer.com/chapter/10.1007/978-3-319-02222-2_15.

Pareja-Lora, A. 2016a. "Enabling Linked-Data-based Semantic Annotations—the Ontological Modeling of Semantics in the OntoLingAnnot Model." In *Proceedings of Term Bases and Linguistic Linked Open Data —TKE 2016, 12th International Conference on Terminology and Knowledge Engineering*, edited by H. Erdman Thomsen, A. Pareja-Lora, and B. Nistrup Madsen, 124–135.

Pareja-Lora, A. 2016b. "Using Ontologies to Interlink Linguistic Annotations and Improve Their Accuracy." In *New Perspectives on Teaching and Working with Languages in the Digital Era*, edited by A. Pareja-Lora, C. Calle-Martínez, and P. Rodríguez-Arancón, 351–362. Dublin: Research-publishing.net.

Pareja-Lora, A., and G. Aguado de Cea. 2010. "Modelling Discourse-related Terminology in OntoLingAnnot's Ontologies." In *Proceedings of the Workshop on Establishing and Using Ontologies as a Basis for Terminological and Knowledge Engineering Resources. TKE 2010: Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges*, 547–574. Dublin, Ireland. http://oa.upm.es/6249/.

Pareja-Lora, A., M. Blume, and B. C. Lust. 2013. "Transforming the DTA Tool Metadata and Labels into a Linguistic Linked Open Data Cloud Resource." In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): "Representing and Linking Lexicons, Terminologies and Other Language Data,"* edited by P. Cimiano, J. McCrae, C. Chiarcos, and T. Declerck, 34–43. Pisa, Italy. September 23, 2013. ACL Anthology, ACL. http://aclweb.org/anthology/W/W13/W13-5506.pdf.

1 Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction

Christian Chiarcos and Antonio Pareja-Lora

Background: Scientific Principles and Openness

In recent decades, the scientific community has become increasingly aware of the importance of openness—for software (open source), publications (open access), structured data (open knowledge), and data collections in general (Open Data). Here, we focus on the latter aspect. Indeed, publishing data collections under open resources has become routine in modern-day research. In this initial chapter, we elaborate on motivations and conventions for publishing Open Data in linguistics and related areas.

The Open Data movement in linguistics—as well as in all areas of study in science, computation, and humanities—draws on three main motivations: (1) responsibility, (2) reproducibility, and (3) reusability.

1. The scientific process—the generation of novel insights, the establishment and revision of paradigms of thought and scientific methodologies, and their documentation, dissemination, and critical reflection—is driven by societal, economic, and ecological need to understand and to develop our past, present, and future. In this sense, scientific research comes with both a privilege and a responsibility: Any projects are supported by public funding, and in return their results should (and in fact are often required to) become available to the public. In the last few decades, this has contributed to the rise of open access in scientific publications, and, along with it, to open source licensing of scientific code and data.
2. Another motivation for the increasing importance of Open Data in research is inherent to the scientific method: Scientific hypotheses must be testable, scientific theories should be verifiable, and published results should be replicable. For data-driven disciplines such as empirical branches of linguistics, verification presupposes the availability of empirical data, while replicability requires access to the original data that the research builds on. Although various distribution and publication models are suitable for this purpose—and have in fact been implemented by agencies such as the Linguistic Data Consortium (LDC) or the European Language Research Association (ELRA); by community portals

such as Perseus,[1] the Cuneiform Digital Library Initiative,[2] and The Language Archive;[3] or within distributed community efforts such as the Universal Dependencies,[4] and UniMorph[5]—publication under an open source license posits the lowest possible barrier for reusability, accessibility, and dissemination of research data.

3. A third practical motivation for publishing (and using) scientific data is the immense effort put into creating such resources and the potential gains of sharing and reusing existing data. In several areas of linguistics, this pertains to primary data, such as recordings, transcripts, and written text; as an extreme example, data collections for languages at the fringe of extinction and/or spoken in remote areas of the world are irreplaceable.

Regardless of the initial motivation, reusability (whether for replication studies, new applications, or novel experiments) is the ultimate goal of publishing Open Data. But secondary reuse of data is not only a concern within linguistics research. It is also an issue relevant to any scientific discipline. In fact, the degree to which an area of research develops and follows agreed-upon principles and standards for the management of data, with respect to its goal of fostering reproducibility, can be regarded as an indicator of its maturity as a scientific discipline.

For linguistics, progress in this direction involves challenges at numerous levels, ranging from political, ethical, and legal issues—for example, community conventions for handling national and international copyright, and privacy issues (for experimental data or field recordings)—to community-wide rules of best practice for documentation, maintenance, and distribution; and beyond those, to the technical question of how to represent, access, and integrate existing data collections.

As a technology, Linked Data allows us to integrate heterogeneous data collections hosted by different data providers, and thus naturally complements the call to Open Data in both science and society. Linked *Open* Data (LOD) describes their conjoint application to a dataset. In application to linguistically relevant datasets, *Linguistic Linked Open Data* (LLOD) describes conventions and a community that has emerged since 2010 whose most prominent outcome is the *Linguistic Linked Open Data cloud* diagram. In this volume, we describe the application of Linked (Open) Data to linguistic data, in particular from the angle of language acquisition.

Open Data in Science

The Open Data movement represents a global change of mind for our understanding of economy, society, and science. In the twenty-first century, a novel paradigm that facilitates both transparency and openness has been emerging. In politics, this has been manifested, for example, in an increased number of Freedom of Information Acts or in the use of Right to Information Laws, among nearly 70 countries in 2006 (Banisar 2006) and more than 100 countries in 2018 (Banisar 2018).

Likewise, the scientific *communis opinio* is increasingly shifting from closed (private) data to Open Data. For its successful implementation, open science does, however, require community standards on how to perform, document, license, and access data publications.

To improve transparency and reproducibility of scientific research, a group of researchers collaborating with M. D. Wilkinson formulated the FAIR Guiding Principles in 2016 (Wilkinson et al. 2016).

F *Findability* implies (1) that data and metadata are assigned globally unique and eternally persistent identifiers, (2) that the data are accompanied by rich metadata, and that (3) the data are registered or indexed in a site where they can be found.

A *Accessibility* implies (1) that data are retrievable by their identifier using an (2) open, free, and universally implemented protocol, and (3) that the protocol supports authentication and authorization if necessary.

I *Interoperability* implies that the data are described using a formal, accessible, shared, and broadly applicable language for knowledge representation.

R *Reusability* implies addition of accurate and relevant attributes, clear licensing and data usage terms and conditions, a linking to provenance of data, and adherence to community standards.

Linked Data represents a technical framework that allows users to tackle these challenges both in general and for the specific needs of linguistics and language technology.

Linked Data

Much of today's data are available in scattered repositories and in diverse formats. In fact, many potentially valuable datasets are being created or shared in data formats intended for human consumption rather than for automated processing. As an example, electronic edition via PDF (Portable Document Format) is still considered state of the art in various disciplines in the humanities; and regularly, spreadsheet or office software is used to create and to fill forms and tables of those PDF documents, without any formal data structures.

Likewise, a popular piece of software in linguistics is optimized for human consumption rather than for machine readability: The Field Linguist's Toolbox[6] provides word- and morpheme-level glossing functionalities. Its underlying format, however, is a plain text format, and the alignment between different layers of morpheme annotation is done by means of whitespaces. However, its current font has an impact on the width of the text displayed, and whitespace alignment between, say, morpheme segmentation and morpheme glossing, or between morpheme segmentation and word segmentation, can only be replicated if the exact widths of each character and each whitespace in the underlying font are known. Unfortunately, many fonts use variable character width, so that, in general, Toolbox segmentation cannot be reliably interpreted or converted into other formats.

These difficulties correspond to problems and needs associated with the Web of Documents in general. First, it is not machine-readable because the data are unstructured. Second, the data are disconnected. Only documents are linked and the meanings of the links are not clear. Third, only a text search is currently feasible.

A proposed solution to these problems is to complement the Web of Documents with the Web of Data, guided by Linked Data principles. The term "Linked Data" was originally published in 2006 as a Design Issue by Tim Berners-Lee (2006) and provides a set of four rules of best practice to be followed for the publication of data on the web. In a slightly reformulated form, these rules are reproduced below.

1. Uniform Resource Identifiers: Use URIs for identifying data and relations.
2. Resolvable via HTTP(S): Use HTTP(S) URIs so that people can look up those names.
3. Standardized formats: For any URI in a dataset, provide useful information using RDF-based standards.
4. Links: Include links to other URIs, so that users can discover more things.

A Uniform Resource Identifier (URI; Berners-Lee et al. 2005) is a compact sequence of characters that identifies an abstract or physical resource. An absolute URI begins with a protocol or a scheme name (e.g., https) followed by an authority (e.g., en.wikipedia.org) and a path (e.g., /wiki/Linguistic_Linked_Open_Data), followed by an optional query (headed by ?) and a fragment (headed by #, e.g., #Linguistic_Linked_Open _Data):

```
https://en.wikipedia.org/wiki/Linguistic _ Linked _ Open _ Data
#Linguistic _ Linked _ Open _ Data
```

This example illustrates that the typical form of a URI in a Linked Data context is a Uniform Resource Locator (URL; Berners-Lee et al. 1994). URLs define a subset of URIs that not only identify a resource, but also provide a means of locating it by describing its primary access mechanism (in this case, the HTTPS protocol). The URI standard is complemented by Internationalized Resource Identifiers (IRIs; Duerst and Suignard 2005), which extend the scope of permissible characters to Unicode: Non-ASCII characters are mapped to ASCII escape sequences by means of the URI percent encoding, as for example the symbol ḡ (Unicode character U+1E21, UTF-8 E1B8A1) as %E1%B8%A1.

The third rule prescribes the use of certain standards. In its original formulation, the standards RDF (data model) and SPARQL (query language) were named. Subsequently, however, additional standards have been developed. Therefore, we interpret this rule nowadays in a way that every data format for which a W3C-standardized interpretation as RDF data exists should be a viable option. This includes native RDF serializations such as Turtle,[7] JSON-LD,[8] or RD F/XML;[9] languages that permit the embedding of RDF content;[10] mapping languages to produce RDF data from other formats;[11] languages that are defined on the basis of RDF;[12] and RDF-based query languages.[13] As data from various sources (CSV files, XML, relational databases, RDF-native data) can be seamlessly converted between different RDF serializations, RDF-based representation formalisms enable data, information consumers, and processors alike to access, interpret, and transform information in a flexible, serialization-independent manner.

The RDF data model formalizes labeled directed multi-graphs, that is, nodes (RDF resources) and relations (RDF properties) that hold between them. Both nodes and relations are identified by means of URIs, and a triple of source node (“subject”), relation (“property”) and target node (“object”) constitutes a statement:

```
<https://en.wikipedia.org/wiki/Linguistic _ Linked _ Open _ Data>
<http://xmlns.com/foaf/spec/primaryTopic>
<http://dbpedia.org/resource/Linguistic _ Linked _ Open _ Data>
. # . marks end of statement, comments after #
```

This example is written in Turtle notation, with whitespace-separated full URIs and . to mark the end of the statement. In addition, Turtle provides a number of practical shorthands, for example the introduction of prefixes. The following Turtle fragment is thus equivalent:

```
PREFIX wpedia: <https://en.wikipedia.org/wiki/>
PREFIX foaf: <http://xmlns.com/foaf/spec/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
wpedia:Linguistic _ Linked _ Open _ Data
foaf:primaryTopic dbpedia:Linguistic _ Linked _ Open _ Data .
```

RDF triples can also take another form, where a source node (“subject”) is assigned a literal value rather than a target node:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
wpedia:Linguistic _ Linked _ Open _ Data
rdfs:label “Linguistic Linked Open Data”@en.
```

Several statements can also be conjoined by means of a semicolon ; (same subject, different property, different object) or a comma (same subject, same property, different object):

```
wpedia:Linguistic _ Linked _ Open _ Data
foaf:primaryTopic dbpedia:Linguistic _ Linked _ Open _ Data ;
rdfs:label “Linguistic Linked Open Data”@en.
```

The fourth rule requires some actual linking, that is, the creation of cross-references between different, distributed datasets, thus enabling a Web of Data to arise along and beside the Web of Documents. This is illustrated in the example above, where a Wikipedia URL and a DBpedia URI are being connected with the RDF property foaf:primaryTopic. The key difference between RDF links and HTML hyperlinks is that the former are semantically typed. Thus, a machine-readable, semantically defined graph representation is created for them, which is not only useful for resource integration on the Web of Data, but also a very generic data structure that finds immediate application in linguistics.

Actually the linking mechanism provides interesting possibilities for scientific datasets, including permitting immediate access to remote datasets and terminology bases. In this way, it becomes possible to share identifiers and to identify concepts and entities corresponding with each other, and thus to harmonize distributed datasets not only on the level of format and means of access, but also on a conceptual level, by means of the use of (or reference to) existing vocabularies. Domain terminology provided in an ontology, for example, can be linked to generic knowledge bases such as the DBpedia,[14] and subsequently enriched with DBpedia information. For instance, assume that we have both a definition of "(technological) singularity" in an English thesaurus and its linking with the English DBpedia:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX my: <http://please.de/fine/by/yourself#>
my:singularity owl:sameAs dbpedia:Technological _ singularity.
```

As the English DBpedia provides a German label, we can immediately return the German labels to our thesaurus concepts and thus apply them to the analysis of another language. This is implemented in the following SPARQL query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?mySingularity ?germanLabel
WHERE {  # for all owl:sameAs links
?mySingularity owl:sameAs ?dbpediaResource.
# find the rdfs labels of the objects
?dbpediaResource rdfs:label ?germanLabel.
FILTER(lang(?germanLabel,'de'))
# and limit the result to German language
}
```

Likewise, large-size databases—of, say, genes, proteins, geographical names, or even movie titles—can be linked over different languages and integrated with each other, so that information from various sources complements each other. There are several reasons for publishing Linked Data: First, it allows ease of discovery through linking. Second, it is easy to consume by both humans and machines. Third, it reduces redundant research and supports collaboration. Fourth, it adds value, visibility, and impact.

Of course, Linked Data is not constrained to *Open* Data, but, obviously, publishing data under open licenses facilitates their accessibility for subsequent adaptation and enrichment. Yet, it is important to remember that not all Linked Data are open and that licensed data can still profit from using standards (enriched with links to Linked Data and/or accessed by standard tools).

Linked Open Data

The definition of Linked Open Data (LOD) is Linked Data that are openly licensed. In 2010, Tim Berners-Lee (Berners-Lee 2006) extended his original Linked Data description with a second component on Open Data. Linked Open Data (LOD) is Linked Data that are released under an open license, such as defined by the Open Definition,[15] where "open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness)."

For promotional reasons, the degree of LOD compliance is expressed by a star scheme, whereby a data publisher receives 1 to 5 stars (*), according to the following requirements:

```
    *    data available as Open Data on the web (e.g., as a scan)
   **    if * using machine-readable, structured format (e.g., DOCX)
  ***    if ** using non-proprietary format (e.g., HTML)
 ****    if *** using open, RDF-based standards
*****    if **** plus linking with other people's data
```

In addition, data publishers are encouraged to publish data along with their metadata and to register these metadata in major catalogs such as http://datahub.io/, or, for linguistic data, in http://linghub.org. From these repositories, the LOD (resp., LLOD) diagrams are being generated.

Linked Open Data has become a trend in scientific research and infrastructures during the 2010s, with prominent resources such as DBpedia (Lehmann et al. 2009), developed within an open-source project with the same name that aimed at extracted structured data from Wikipedia and related resources. DBpedia version 2016–10 includes extractions in 134 languages with a total of over 13 billion RDF statements (triples). With more and more datasets being linked with DBpedia and other LOD datasets, a Linked Open Data cloud has emerged, and as a visualization of the growing Web of *Open* Data, this process has been documented with a series of LOD cloud diagrams.[16] As of October 2018, the diagram contained 1,229 datasets with 16,125 links (figure 1.1). Primary applications of RDF technology and LOD resources are concerned with resource integration and also with resource reuse. Hence, major components of the LOD cloud diagram are term bases such as statistical government data, or biomedical databases, and indeed the key advantage of RDF technology and LOD resources is their high level of reusability and accessibility. SPARQL 1.1 supports the concept of federation: By means of the SERVICE keyword, it is possible to consult external SPARQL endpoints (RDF databases with web interfaces) as part of a query against a local triple (or quad) store.

In fact, resources can be freely shared and cloned, and redundant copies can contribute to the sustainability of LOD datasets independently from the institution that originally provided those data or their technical infrastructures.

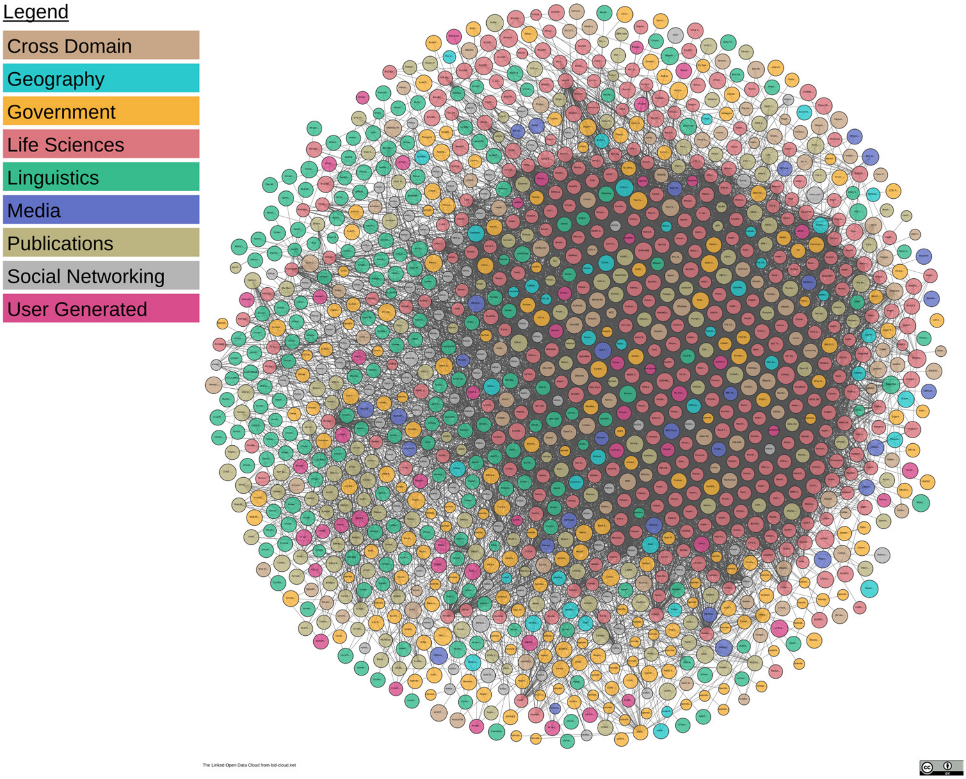


Figure 1.1
LOD diagram, version of Oct. 31, 2018, https://lod-cloud.net.

Indeed, such redundant copies are normally created as LOD are processed: While the SPARQL 1.1 protocol allows users to access remote SPARQL endpoints by means of the SERVICE keyword, and remote RDF dumps by means of LOAD, both come with a certain degree of overhead, and thus lead to runtime reductions for applications that consume Linked Data. Real-world applications of LOD thus normally work on local copies, instead, so that redundant and distributed copies are created as a side effect of LOD-based applications.

For scientific applications, another factor of LOD is important—that is, that different applications can refer to the same term in the same database. Thus results, data, and annotations can all be traced over different datasets while information about them can be put in relation with each other. Of course, the same applies, even to a larger extent, to vocabularies used for different resources. With increased reusability and reuse of scientific datasets, the datasets serve as models for the vocabulary of subsequent resources, and indeed research in community-based vocabulary development has intensified in recent years.

We also must admit that LOD comes with a number of technical challenges. LOD and RDF technology both provide a high-level view as well as a generic technology for processing and integrating different data sources, and of course "genericity" does come with a price. The potential of RDF and RDF-based technology in comparison to classical relational databases can thus be compared to the gains and challenges of high-level programming languages (such as Java or Python) in comparison to low-level programming languages (such as machine code or assembler). However, for many problems, processing RDF (cf. Python) will be considerably slower than using an implementation-specific SQL dialect (cf. assembler), though it excels in portability, reusability, and development effort. In particular, RDF is superior at dealing with sparse and heterogeneous data, but for densely populated databases, RDF technology is slow in comparison with classical relational database technology. Unlike SQL, RDF technology allows users to reach out beyond a data silo and to seamlessly link data with external resources.

One specific challenge in this context is that links between resources and resources themselves were created for different purposes, according to different methodologies and are maintained by different providers. This can lead to inconsistencies in the interpretation and in the quality of statements (triples) they provide. An increasingly important aspect is thus the tracing of provenance and related metadata, so that scientific and industry applications alike can (and should) inspect the composition of data aggregated from LOD and must not blindly rely on their correctness.

In summary, Linked Open Data is enabling a change of data and information readers and processors in that it enables us to abstract from resource-specific formats and representations and technologies, and then to integrate information over distributed datasets. Linked *Open* Data represents the core of the emerging Web of Data and thus enables a global change of data and information management and processing. LOD comes with rich technological support, in terms of portable means of access and representation (W3C-standardized data models, formats, protocols, and query languages), in terms of technical support with off-the-shelf databases, and in terms of the existence of a considerable developer and user community. At the same time, many scientific challenges in relation to LOD core techniques seem to have been solved, so that the focus in LOD research has moved from foundations and basic standards to applications. A recent development in this regard is the publication of domain-specific sub-clouds, which since August 2018 have been available as LOD addenda diagrams. Linguistic Linked Open Data represents one such area of application.

Linked Open Data in Linguistics

As is true of any field of scientific research, the FAIR principles are relevant for linguistics, language studies, and natural language processing—that is, for the digital language resources they produce and build on—and indeed Bird and Simons (2003) formulated comparable requirements and best practice recommendations for language resources 15 years ago, which we have reorganized and slightly reworded below according to the FAIR principles.

As far as technical and legal aspects are concerned, RDF and (Linguistic) Linked (Open) Data provide an ideal framework to implement these requirements. In the enumeration below, this is illustrated with a ± ranking ranging from − to +++.[17]

F findability

existence at a data provider++ : Register language resources at a major resource portal. In a Linguistic Linked Open Data context, this would be LingHub (http://linghub.org/) or one of the resource portals it builds on.[18]

relevance/discovery+ : Provide metadata according to community-approved conventions and vocabularies.

persistence+ : Provide persistent identifiers to language resources (e.g., a persistent URL) and unique identifiers for components of a language resource.

long-term preservation+ : Provide long-term preservation by hosting at an institution committed to that purpose.

A accessibility

open format+++ : Provide data in an open format supported by multiple tools.

complete access+ : Provide direct access to the full data and documentation.

unimpeded access+++ : Provide documentation about the methods of access.

universal access+++ : Provide universal access to every interested user.

I interoperability

terminology++ : Map linguistic terms and markup elements to a common ontology.

format documentation+++ : Provide data in a self-describing format (including XML, RDF, JSON).

machine-readable format+++ : Use open standards such as those provided by the W3C (Unicode, XML, etc.).

human-readable format+ : Provide human-readable versions of the material.

R reusability

rich content++ :[19] Provide rich and linguistically relevant content.

accountability+ : Fully document both the resource and its source data.

provenance+ : Provide provenance and attribution metadata.

immutability+ : Provide immutable, fixed versions of a resource, with appropriate versioning.

legal documentation+++ : Document intellectual property rights of all components of the language resource.

research license+++ : Ensure that the resource may be used for research purposes.

complete preservation+++ : Make sure that all aspects of the language resource and its documentation remain accessible in the future (i.e., independent from any particular software).

Current accessibility challenges arise in the different formats and schemes of documents, their distribution, and the dispersed nature of metadata collections. There have long been efforts to recognize and address these problems, but these activities were never coordinated. In particular, RDF was used, but resources were rarely linked to other resources in the Web of Data. So a community needed to be built. Since 2010, the increasing popularity of applying RDF to language resources and the potential for creating links between different datasets led (1) to the formation of the Open Linguistics Working Group of Open Knowledge International[20] and, subsequently (2) to the emergence of a Linguistic Linked Open Data (LLOD) cloud, as well as (3) to the development of community conventions for the publication of linguistically relevant datasets on the Web of Data.

Open Knowledge International is a nonprofit organization, founded in 2004, that promotes open knowledge in all its forms (e.g., publication of government data in the UK and USA); it provides infrastructural support for several working groups. The Open Linguistics Working Group of the Open Knowledge Foundation (OWLG) was organized in October 2010 in Berlin, Germany, and assembled a network of individuals interested in linguistic resources and/or their publication under open licenses. The OWLG is multidisciplinary and has infrastructure in the forms of a mailing list and a website.[21] Its most important activities are the organization of community events such as workshops, datathons/summer schools and conferences, and the ongoing development of the Linguistic Linked Open Data (sub-) cloud, currently maintained under http://linguistic-lod.org/.

The Linguistic Linked Open Data (LLOD, figure 1.2) cloud is a collection of linguistic resources that have been published under open licenses as Linked Data. It is decentralized in its development and maintenance and was developed as a community effort in the context of the Open Linguistics Working Group of the Open Knowledge Foundation. Initially, the OWLG maintained a list of open or representative resources; in January 2011, this group marked possible synergies between these resources in the first draft of a LLOD cloud diagram. At this time, it was merely a vision, and the draft included non-open resources as placeholders for other resources to come, though none have been realized. In the closing chapter of their contributed volume on Linked Data in Linguistics, Chiarcos, Nordhoff, and Hellmann (2012) provided a hypothetical linking for selected datasets from NLP, Semantic Web, and linguistic typology described in the book. In September 2012, the

LLOD cloud diagram materialized as a result of the first datathon on Multilingual Linked Open Data for Enterprises (MLODE-2012). Since 2012, more data and more rigid quality constraints have been added, collaborations with national and international research projects have been established, and related W3C community groups have emerged.

With the increasing popularity of LLOD, in August 2014 "linguistics" was recognized as a top-level category of the colored LOD cloud diagram, with LLOD resources formerly having been classified into other categories. In August 2018, a copy of the LLOD cloud diagram was incorporated into the LOD cloud diagram as a domain-specific addendum. Within the LOD cloud, Linguistic Linked Open Data is growing at a relatively high rate. While the annual growth of the LOD cloud (in terms of new resources added) over the last two years has been at 10.2% on average for the LOD cloud diagram, the LLOD cloud diagram itself has been growing at 19.3% per year (cf. figure 1.3).

Aside from its maintaining the LLOD cloud diagram, the OWLG aims to promote open linguistic resources by raising awareness and collecting metadata, and aims to facilitate a wide range of community activities by hosting workshops, using its extensive mailing list, and

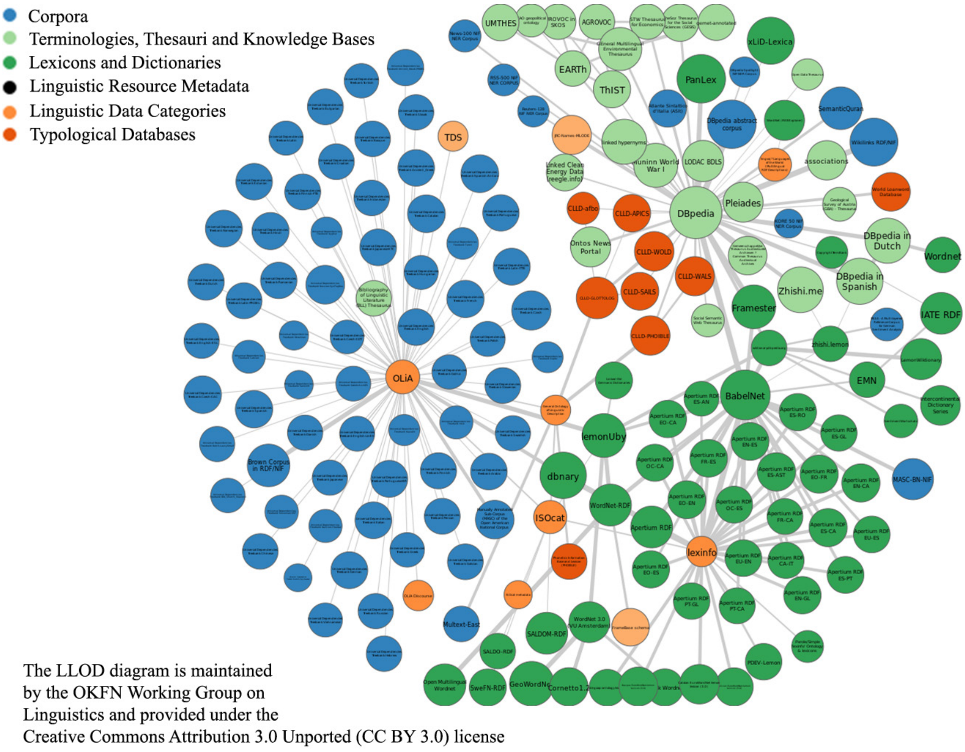


Figure 1.2
Linguistic Linked Open Data (LLOD) cloud diagram, version of August 2017, http://linguistic-lod.org.

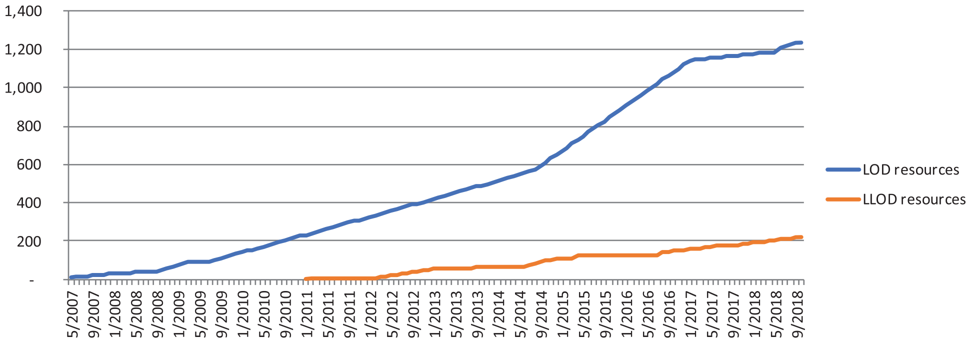


Figure 1.3
Number of resources in the LOD and LLOD cloud diagrams, corresponding, respectively, to the periods 2007–2018 and 2011–2018.

creating various publications. In doing so, they facilitate exchange between and among more specialized community groups, such as the W3C community groups (for instance, the Ontology-Lexica Community Group (OntoLex),[22] the Linked Data for Technology Working Group [LD4LT],[23] or the Best Practices for Multilingual Linked Open Data Community Group [BPM- LOD]).[24]

At the time of writing, the most vibrant of these W3C community groups is the OntoLex group, which is developing specifications for lexical data in a LOD context; this need correlates with the high popularity among LLOD resources of the OntoLex vocabulary (Cimiano, McCrae, and Buitelaar 2016). Whereas specifications for lexical resources are relatively mature, as are term bases for either language varieties (de Melo 2015; Nordhoff and Hammarström 2011) or linguistic terminology (Aguado de Cea, Álvarez de Mon, Gómez-Pérez, and Pareja-Lora 2004; Chiarcos 2008; Chiarcos and Sukhareva 2015), the process of developing widely applied data models for other types of language resources, such as corpora and data collections in general, is still ongoing. To a certain extent, this volume aims to contribute to this discussion and its future development.

Chances, Challenges, and Prospects

The individual contributions herein document progress made in the field of Linguistic Linked Open Data since 2012 (Chiarcos et al. 2012). One important difference in comparison to developments in that year—a time when the community was largely building on small-scale experiments and imagining a bright vision of the future—is that providers of existing infrastructures and of existing platforms are increasingly getting involved in both the process and the discussion; this is reflected by the contributors to this volume.

The general situation is that a remarkable amount of Linguistic Linked Open Data is already available, an amount that continues to steadily grow. In a longer perspective, we can expect additional data providers to offer an L(O)D view on their data, and to support RDF serializations such as JSON-LD as interchange formats. However, LOD's further growth and popularity depend crucially on the development of applications that are capable of either consuming these data in a linguist-friendly fashion or of enriching local data with wide-ranging web resources.

At the time of writing, working with RDF normally requires a certain level of technical expertise—at minimum, basic knowledge of SPARQL and of at least one RDF format. The authors' personal experience in teaching university courses shows that linguists *can* be successfully trained to acquire both. However, this is not normally done and is unlikely to ever be part of the linguistics core curriculum. This may change once designated textbooks on Linked Open Data for NLP and for linguistics become available, but for the time being a priority for this effort and the wider community remains to provide concrete applications tailored to the needs of linguists, lexicographers, researchers in NLP, and knowledge engineers.

Promising approaches in this direction do exist: Existing tools can be complemented with an RDF layer to facilitate their interoperability. This is the scope of several chapters in this volume. Likewise, LLOD-native applications are possible—for instance, to use RDFa (RDF in attributes; Herman et al. 2015) to complement an XML workflow with SPARQL-based semantic search by means of web services (Tittel et al. 2018); to provide aggregation, enrichment, and search routines for language resource metadata (Chiarcos et al. 2016; McCrae and Cimiano 2015); to use RDF as a formalism for annotation integration and data management (Burchardt et al. 2008; Pareja-Lora 2012; Chiarcos et al. 2017); or to use RDF and SPARQL for manipulating and evaluating linguistic annotations (Chiarcos, Khait et al. 2018; Chiarcos, Kosmehl et al. 2018). While these applications demonstrate the potential of LOD technology in linguistics, they come with a considerable entry barrier, and they address the advanced user of RDF technology rather than a typical linguist. Even though concrete applications do exist, the path remains long to reaching the level of user-friendliness that occasional users of this technology might expect.

A notable exception in this regard is LexO (Bellandi, Giovannetti, and Piccini 2018), a graphical tool for collaboratively editing lexical and ontological resources that natively build on the OntoLex vocabulary and RDF; LexO was designed to conduct lexicographical work in a philological context (for instance, creating the *Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan*). Other projects whose objective is to provide LLOD-based tools for specific areas of application have been recently approved, so progress in this direction is happily to be expected within the next years.[25]

Acknowledgments

This chapter originates from a joint presentation given by Antonio Pareja-Lora, Martin Brümmer, and Christian Chiarcos at the 2015 LSA workshop titled "Development of Linguistic Linked Open Data (LLOD) Resources for Collaborative Data-Intensive Research in the Language Sciences." On the one hand, the work of the first author has been partially supported by the German Federal Ministry for Science and Education (BMBF) in the context of the Research Group *Linked Open Dictionaries* (LiODi, 2015–2020). On the other hand, the work of the second author has been partially supported by the projects RedR+Human (Dynamically Reconfigurable Educational Repositories in the Humanities, ref. TIN2014-52010-R) and CetrO+Spec (Creation, Exploration and Transformation of Educational Object Repositories in Specialized Domains, ref. TIN2017-88092-R), both financed by the Spanish Ministry of Economy and Competitiveness.

Notes

1. Greek and Latin literature, http://www.perseus.tufts.edu.

2. Ancient Mesopotamian philology, https://cdli.ucla.edu.

3. Data archive about languages worldwide, https://tla.mpi.nl/.

4. Cross-linguistically comparable syntax annotations, https://universaldependencies.org/.

5. Cross-linguistically comparable morpheme inventories, http://unimorph.github.io/.

6. https://software.sil.org/toolbox/.

7. https://www.w3.org/TR/turtle/.

8. https://www.w3.org/TR/json-ld/.

9. https://www.w3.org/TR/rdf-syntax-grammar/.

10. This includes HTML+RDFa (https://www.w3.org/TR/html-rdfa/), XHTML+RDFa (https://www.w3.org/TR/xhtml-rdfa/), or XML+RDFa (https://www.w3.org/TR/rdfa-core/).

11. Using standards such as CSV2RDF (https://www.w3.org/TR/csv2rdf/), the RDB to RDF Mapping language R2RML (https://www.w3.org/TR/r2rml/), or the Direct Mapping of Relational Data to RDF (https://www.w3.org/TR/rdb-direct-mapping/).

12. Including the Web Ontology Language OWL (https://www.w3.org/TR/2012/REC-owl2-mapping-to-rdf-20121211/) or the Simple Knowledge Organization System SKOS (https://www.w3.org/2009/08/skos-reference/skos.html).

13. For example, SPARQL (https://www.w3.org/TR/sparql11-query/) or SHACL (https://www.w3.org/TR/shacl/).

14. https://wiki.dbpedia.org/.

15. The Open Definition and compliant licenses can be found under http://opendefinition.org.

16. Available under https://lod-cloud.net/.

17. Ranking criteria and number of Bird and Simons requirements per category.

- impossible with LOD 0/19

+ possible with/encouraged by LOD, but not required 8/19

++ required by LOD 3/19

+++ required in a more specific or stricter form by (L)LOD 8/19

18. http://datahub.io/, https://vlo.clarin.eu, http://metashare.elda.org/; for language documentation data, the Open Language Archives Community (OLAC, http://www.language-archives.org/) would be an option; it provides an RDF dump, but its metadata are not yet imported into LingHub.

19. Linguistic relevance is a requirement for Linguistic Linked Open Data, but of course not for LOD data.

20. https://linguistics.okfn.org/.

21. https://linguistics.okfn.org/, https://lists.okfn.org/mailman/listinfo/open-linguistics.

22. https://www.w3.org/community/ontolex.

23. https://www.w3.org/community/ld4lt/.

24. https://www.w3.org/community/bpmlod.

25. This includes, for example, the projects POSTDATA (on European poetry, 2015–2020, funded by the European Research Council), Linked Open Dictionaries (on language contact studies, 2015–2020, funded by the German Federal Ministry of Education and Science), Linking Latin (on Latin philology, 2018–2023, funded by the European Research Council), and the Horizon 2020 Research and Innovation Action Prêt-à-LLOD (2019–2021).

References

Aguado de Cea, G., I. Álvarez de Mon, A. Gómez-Pérez, and A. Pareja-Lora. 2004. "OntoTag's Linguistic Ontologies: Improving Semantic Web Annotations for a Better Language Understanding in Machines." In Proceedings of the International Conference on Information Technology: Coding and Computing, 2004 (ITCC 2004), Vol. 2, 124–128. Las Vegas, Nevada, USA.

Banisar, D. 2006. "Freedom of Information around the World 2006: A Global Survey of Access to Government Information Laws." Technical report, Privacy International. Version of September 20, 2006.

Banisar, D. 2018. "National Right to Information Laws, Regulations and Initiatives 2018." Technical report, Privacy International. Version of September 21, 2018.

Bellandi, A., E. Giovannetti, and S. Piccini. 2018. "Collaborative Editing of Lexical and Termino-ontological resources: A Quick Introduction to LexO." In the XVIII EURALEX International Congress. Lexicography in Global Contexts, 23–27. Ljubljana, Slovenia.

Berners-Lee, T. 2006. Design issues: Linked Data. http://www.w3.org/DesignIssues/LinkedData.html. Revised 2010.

Berners-Lee, T., R. Fielding, and L. Masinter. 2005. Request for Comments: 3986. Uniform Resource Identifier (URI): Generic syntax. Technical report, The Internet Society. Network Working Group. Version of January 2005.

Berners-Lee, T., L. Masinter, and M. McCahill. 1994. Request for Comments: 1738. Uniform Resource Locators (URL). Technical report, Internet Engineering Task Force (IETF). Network Working Group. Version of December 1994.

Bird, S., and G. Simons. 2003. "Seven Dimensions of Portability for Language Documentation and Description." *Language* 79:557–582.

Burchardt, A., S. Padó, D. Spohr, A. Frank, and U. Heid. 2008. "Formalising Multi-layer Corpora in OWL/DL—Lexicon Modelling, Querying and Consistency Control." In *Proceedings of the 3rd International Joint Conf on NLP* (IJCNLP 2008), 389–396. Hyderabad, India.

Chiarcos, C. 2008. "An Ontology of Linguistic Annotations." *LDV Forum* 23 (1): 1–16.

Chiarcos, C., C. Fäth, H. Renner-Westermann, F. Abromeit, and V. Dimitrova. 2016. "Lin|gu|is|tik: Building the Linguist's Pathway to Bibliographies, Libraries, Language Resources and Linked

Open Data." In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA).

Chiarcos, C., M. Ionov, M. Rind-Pawlowski, C. Fäth, J. W. Schreur, and I. Nevskaya. 2017. "LLOD-ifying Linguistic Glosses." In International Conference on Language, Data and Knowledge (LDK 2017), edited by J. Gracia, F. Bond, J. McCrae, P. Buitelaar, C. Chiarcos, and S. Hellmann, 89–103. Galway, Ireland. Cham: Springer. Lecture Notes in Computer Science, vol. 10318.

Chiarcos, C., I. Khait, É. Pagé-Perron, N. Schenk, C. Fäth, J. Steuer, W. Mcgrath, J. Wang, et al. 2018. "Annotating a Low-Resource Language with LLOD Technology: Sumerian Morphology and Syntax." *Information* 9 (11): 290.

Chiarcos, C., B. Kosmehl, C. Fäth, and M. Sukhareva. 2018. "Analyzing Middle High German Syntax with RDF and SPARQL." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). Miyazaki, Japan, ELRA.

Chiarcos, C., S. Nordhoff, and S. Hellmann. 2012. "Linked Data in Linguistics." Berlin/Heidelberg: Springer.

Chiarcos, C. and M. Sukhareva. 2015. "OLiA—Ontologies of Linguistic Annotation." *Semantic Web Journal* 518:379–386.

Cimiano, P., J. McCrae, and P. Buitelaar. 2016. "Lexicon Model for Ontologies." Technical report, W3C Community Report, May 10, 2016.

de Melo, G. 2015. "Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud." *Semantic Web Journal* 6 (4): 393–400.

Duerst, M. and M. Suignard. 2005. Request for Comments: 3987. Internationalized Resource Identifiers (IRIs). Technical report, The Internet Society. Network Working Group. version of January 2005.

Herman, I., B. Adida, M. Sporny, and M. Birbeck. 2015. "RDFa 1.1 Primer." 3d ed. W3C working group note, World Wide Web Consortium.

Lehmann, J., C. Bizer, G. Kobilarov, et al. 2009. "DBpedia—A Crystallization Point for the Web of Data." *Journal of Web Semantics* 7 (3): 154–165.

McCrae, J. P., and P. Cimiano. 2015. "Linghub: A Linked Data-based Portal Supporting the Discovery of Language Resources." In Proceedings of the 11th International Conference on Semantic Systems (SEMANTiCS 2015), 88–91.Vienna, Austria.

Nordhoff, S., and H. Hammarström. 2011. "Glottolog/Langdoc: Defining Dialects, Languages, and Language Families as Collections of Resources." In First International Workshop on Linked Science (LISC-2011), held in conjunction with ISWC 2011. Bonn, Germany.

Pareja-Lora, A. 2012. *Providing Linked Linguistic and Semantic Web Annotations: The OntoTag Hybrid Annotation Model*. Saarbrücken: LAP–LAMBERT Academic Publishing.

Tittel, S., H. Bermúdez-Sabel, and C. Chiarcos. 2018. "Using RDFa to Link Text and Dictionary Data for Medieval French." In Proceedings of the Sixth Workshop on Linked Data in Linguistics (LDL-2018), 7–12. Miyazaki, Japan, ELRA.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (160018).

2 Whither GOLD?

D. Terence Langendoen

In the Beginning

In the Language Digitization Workshop, the kickoff meeting for the Electronic Metadata for Endangered Language Data (E-MELD) project,[1] in Santa Barbara, California, in June 2001, I made two presentations on linguistic markup (annotation). The first described the general nature of the markup of sound and text files and of databanks that can be derived from them, and the second the work of the Text Encoding Initiative (TEI)[2] on text markup, particularly the chapters on simple analytic mechanisms (SAM), feature structures (FS), and feature system declarations (FSD) of Sperberg-McQueen and Burnard, editors (1994).[3] In these presentations, I made the following observations.

1. For electronically encoded resources to be maximally useful within and across linguistic communities, there must be agreement on transcription and analytic terminology standards within and across languages, with procedures for settling differences among transcription and terminological practices.
2. Linguistic databanks can be developed along the lines of commonly used print data resources, such as comparative wordlists, morphosyntactic paradigms, thesauri, rhyming dictionaries, mono- and multilingual sense dictionaries, and reference grammars, in addition to digitally born types of databanks, such as treebanks and interlinear glossed text (IGT) repositories.
3. Because the TEI recommendations for FS and FSD have not been widely adopted, presumably because of their complexity and the lack of extensive testing on linguistic data,[4] it might be a good idea to try to reach consensus on a simpler form of FS markup using XML that would be adequate to the needs of the linguistics community.[5]

The Birth of GOLD

However, my two newly recruited research assistants, Scott Farrar and Will Lewis, quickly convinced me that a better path would be to take advantage of the infrastructure of the Semantic Web announced in Berners-Lee, Hendler, and Lassila (2001) that was under development as a reasoning platform for all publicly shared data on the web. Specifically, we would begin to build an ontology for the concepts needed for linguistic analysis as a subcomponent of an upper ontology, such as of SUMO, the Standard Upper Merged Ontology (Pease and Niles 2002; Pease 2007). This ontology, like SUMO and like other domain-specific web ontologies, would be written in one of the markup languages being constructed for the Semantic Web, such as OWL-DL, not in XML. Technically, this did not violate the E-MELD project's endorsement of XML as the markup language of choice for linguistic annotation. Such annotation could still be written in XML but the interpretation of its tags would be determined by the concepts they referred to (i.e., pointed to) in GOLD. FS would be treated as a data type, with its interpretation determined by its connections to GOLD. In Lewis, Langendoen, and Farrar (2001), our first presentation following the kickoff meeting, we pointed out that the real need of the community we were serving would be "to obtain information about endangered languages on the World Wide Web without regard to the tagging schemes that are used in the various websites they consult. Thus [we] cannot impose a markup standard for endangered language websites, even implicitly by developing a data interchange format [such as the TEI]." To make sense of this markup chaos, we proposed the development of a "metatagging" scheme consisting of "a knowledge base and its accompanying tools [that] will act as an interlingua for data comparison," the key to which is an ontology. At the time we submitted the paper for presentation, we had already created an ontology for morphosyntactic concepts with hundreds of nodes drawn from resources provided by the Summer Institute of Linguistics and the Dokumentation Bedrohter Sprachen (DOBES) project, two general linguistics term sets and several dictionaries and grammars of endangered languages, but we had not yet given it a name. At the workshop, we announced our choice: the General Ontology for Linguistic Description (GOLD).

The Development of GOLD within the E-MELD Project

Presentations about GOLD were made at every annual E-MELD workshop from 2002 through the end of the project in 2006, as well as at numerous conferences and workshops around the world, including Langendoen, Farrar, and Lewis (2002), Farrar, Lewis, and Langendoen (2002), Farrar and Langendoen (2004), Simons et al. (2004b), and Lewis (2006). GOLD came to the attention of the linguistics community at large through the publication of Farrar and Langendoen (2003), and the Linguist List began hosting GOLD's website in 2006.[6] Two major accomplishments occurred during this period. First, a proof of concept was achieved for the metatagging scheme proposed in Lewis, Langendoen, and Farrar (2001) to carry out searches over differently encoded datasets of IGT and electronic dictionaries (Simons et al. 2004a, 2004b). Second, the Online Database of Interlinear Text (ODIN) was set up, in which users could select from a list of GOLD morphosyntactic concepts and find instances of IGT harvested from the web in more than 700 languages that contain morphemes referencing them (Lewis 2006). However, little other progress was made beyond the further refinements of the conceptual structure for morphosyntax, a situation that has continued to this day.

GOLD after E-MELD

At the conclusion of the E-MELD project in 2006, Scott Farrar continued his work for several more years on GOLD's conceptual backbone, particularly on the notion of the linguistic sign itself (Farrar 2007), and on the relative merits of the various versions of OWL for implementing GOLD (Farrar and Langendoen 2010). Will Lewis along with Fei Xia and other collaborators have extended the ODIN's data coverage to nearly 1,300 languages and over 130,000 instances (Lewis and Xia 2010; Xia et al. 2014).[7] Finally, the Lexical Enhancement via the GOLD Ontology (LEGO) project—begun in 2008 under the direction of two of the E-MELD principal investigators, Anthony Aristar and Helen Aristar-Dry, together with Jeff Good—has tagged the entries of 12 lexicons and 11 wordlists with links to GOLD concepts to support cross-linguistic search much in the manner of ODIN.[8] Neither project, however, has extended GOLD's conceptual coverage.

What's Next?

The question of how to sustain the GOLD effort at the end of the E-MELD project was considered by Farrar and Lewis (2007), who proposed that communities of practice take responsibility for constructing GOLD subcomponents for particular languages and language families, and collaborate on determining which cross-linguistic constructs should be incorporated into GOLD itself. However, no effective action has yet been taken on their recommendations. Bender and Langendoen (2010: sec. 4) envisioned a future research environment for linguists called Digital Infrastructure that supports Linguistic Inquiry (DILI) that builds on past and current work and provides the following three capacities, among others:

1. Ready access to large amounts of digital data in text, audio, and audio-video media about many languages, which are relevant to many different areas of research and application both within and outside of linguistics.
2. Facilities for comparing, combining, and analyzing data across media, languages, and subdisciplines, and efforts to enrich DILI with their results.
3. Services to support seamless collaboration across space, time, (sub)disciplines, and theoretical perspectives.

We went on to say, "It is not required that there be a single overarching network for all the annotations in DILI, but it would be desirable if sense could be made of the relations among conceptual networks for different annotation schemes, particularly those that represent different theoretical perspectives.… This view of the role of conceptual encoding was recently articulated in Farrar and Lewis (200[7]), along with a plan for how to achieve it." Lest this vision be dismissed as pie-in-the-sky fantasy, we pointed out that similarly ambitious research environments already exist for such fields as biochemistry, nanotechnology, and astronomy—so why not linguistics?

Perhaps the lack of such research environments in linguistics is a result of the long history of our field, which sprang up independently in varying language and cultural communities in several parts of the world, or perhaps it's the fractiousness of us linguists, or even the notion that it's harder for ours than for most, if not all, others' fields of inquiry. I think of Scott Farrar, struggling with the problem of characterizing the notion of the linguistic sign for use in GOLD, who finally formulated something that came fairly close to what Louis Hjelmslev (1943 [1962]) proposed. If Farrar is at least in the right ballpark, then the underlying logic will have to be richer than that provided by OWL-DL, which is a decidable version of first-order logic, even putting aside the wondrous complexities of the logical forms needed to represent, for example, reciprocal constructions in the world's languages.[9] The reason is that Farrar's forms have to relate to each other compositionally, both for meaning and for expression.[10] The composition of meanings is governed by whatever conceptual (logical) operation is called for to combine them, such as binding a predicate variable by a quantifier. At the same time, the composition of expressions is governed by a mereological (also logical, but with a different partial ordering) operation such as concatenation, if the expressions are represented as strings, so that at least two distinct logical systems have to be synchronized. The challenge, I think, is well worth undertaking, starting with our taking a fresh look at the proper way to construct conceptual networks for linguistic analysis and annotation.

Notes

1. E-MELD was funded by the US National Science Foundation grant 0094934 to Wayne State University with a subcontract to the University of Arizona.

2. TEI was funded by the US National Endowment for the Humanities, Directorate General XIII of the Commission of the European Communities, Andrew W. Mellon Foundation, and Social Science and Humanities Research Council of Canada.

3. As chair of the TEI Committee on Text Analysis and Interpretation and of the Work Group on Linguistic Description, I had overall responsibility for the preparation of these chapters. The editors and the members of the committee and of the work group were active contributors, particularly Mitch Marcus, who convincingly argued for the importance of FS at the first work group meeting, and Gary Simons, who showed how sets of FS can be validated by FSD, the latter being in effect (partial) grammars of the languages described by those FS sets; see Langendoen and Simons (1995).

4. Mitch Marcus and I gave a tutorial entitled "Tagging Linguistic Information in a Text Corpus" at the June 1990 ACL meeting in Pittsburgh, in which we described the guidelines in preparation for both the Penn Treebank (PTB) and the TEI recommendations for SAM and FS. The PTB, along

with its encoding scheme for English syntactic structure, eventually caught on to become a major resource for computational linguists; the TEI recommendations did not. I still vividly recall Ken Church's making exactly that prediction following our presentation.

5. At its kickoff meeting, the E-MELD project endorsed XML as the markup language it would recommend for linguistic annotation. TEI was originally encoded in SGML but later converted to XML.

6. http://linguistics-ontology.org/.

7. http://odin.linguistlist.org and http://faculty.washington.edu/fxia/odin/. More recently, the ODIN resource has been enriched with the addition of syntactic tiers, and graphical interface tools, but with the links to GOLD removed. For details see Xia et al. (2016).

8. LEGO was supported by the US National Science Foundation award 0753321 to Eastern Michigan University; see http://lego.linguistlist.org.

9. Berners-Lee, Hendler, and Lassila (2001) insisted that the Semantic Web should not deal with the semantics of natural languages. Still, the conceptual networks for linguistic annotation *will* eventually have to deal with them.

10. And for other things as well, but I leave them also aside.

References

Bender, Emily M., and D. Terence Langendoen. 2010. "Computational Linguistics in Support of Linguistic Theory." *Linguistic Issues in Language Technology* 3 (1): 1–31.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284 (5).

Farrar, Scott. 2007. "Using 'Ontolinguistics' for language description." In *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*. Edited by Andrea Schalley and Dietmar Zaefferer, 175–192. Berlin: Mouton de Gruyter.

Farrar, Scott, and D. Terence Langendoen. 2003. "A Linguistic Ontology for the Semantic Web." *Glot International* 7 (3): 97–100.

Farrar, Scott, and D. Terence Langendoen. 2004. "Comparability of Language Data and Analysis: Using an Ontology for Linguistics." *Symposium on Endangered Data vs. Enduring Practice, 80th Annual Meeting of the Linguistic Society of America*, Boston.

Farrar, Scott, and D. Terence Langendoen. 2010. "An OWL-DL Implementation of GOLD: An Ontology for the Semantic Web." *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology.* Edited by Andreas Witt and Dieter Metzing. Dordrecht, Netherlands: Springer.

Farrar, Scott, and William D. Lewis. 2007. "The GOLD Community of Practice: An Infrastructure for Linguistic Data on the Web." *Language Resources and Evaluation* 41 (1): 45–60.

Farrar, Scott, William D. Lewis, and D. Terence Langendoen. 2002. "An Ontology for Linguistic Annotation." *Semantic Web Meets Language Resources: Papers from the AAAI Workshop* (Technical Report WS-02–16), 11–19. Menlo Park, CA: AAAI Press.

Hjelmslev, Louis. 1962. *Prolegomena to a Theory of Language*, 2d rev. Translation by Frances J. Whitfield. Madison: University of Wisconsin Press. Originally published in 1943 as *Omkring Sprogteoriens Grundlæggelse*. Copenhagen: Munksgaard; reprinted 1966. Copenhagen: Akademisk Forlag.

Langendoen, D. Terence, Scott Farrar, and William D. Lewis. 2002. "Bridging the Markup Gap: Smart Search Engines for Language Researchers." *Proceedings of the Workshop on Resources and Tools for Field Linguistics*, *Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands.

Langendoen, D. Terence, and Gary F. Simons. 1995. "A Rationale for the Text Encoding Initiative Recommendations for Feature-Structure Markup." *Computers and the Humanities* 29:191–205. Reprinted in *The Text Encoding Initiative: Background and Context*, edited by Nancy Ide and Jean Veronis, 191–210. Dordrecht, Netherlands: Kluwer.

Lewis, William D. 2006. "ODIN: A Model for Adapting and Enriching Legacy Infrastructure." *Proceedings of the E-Humanities Workshop Held in Cooperation with E-Science 2006: 2nd IEEE International Conference on E-Science and Grid Computing.* Amsterdam.

Lewis, William D., D. Terence Langendoen, and Scott Farrar. 2001. "Building a Knowledge Base of Morphosyntactic Terminology." *Proceedings of the IRCS Workshop on Linguistic Databases,* 150–156. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania.

Lewis, William D., and Fei Xia. 2010. "Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages." *Journal of Literary and Linguistic Computing* 25 (3): 303–319.

Pease, Adam. 2007. "Formal Representation of Concepts: The Suggested Upper Merged Ontology and Its Use in Linguistics." In *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*, edited by Andrea Schalley and Dietmar Zaefferer, 103–114. Berlin: Mouton de Gruyter.

Pease, Adam, and Ian Niles. 2002. "Towards a Standard Upper Ontology: A Progress Report." *Knowledge Engineering Review* 17 (1): 65–70.

Schalley, Andrea, and Dietmar Zaefferer, eds. 2007. *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*. Berlin: Mouton de Gruyter.

Simons, Gary F., Brian Fitzsimons, D. Terence Langendoen, William D. Lewis, Scott Farrar, Alexis Lanham, Ruby Basham, et al. 2004a. "A Model for Interoperability: XML Documents as a Distributed Database." *E-MELD Workshop on Databases for Field Linguistics*, Detroit.

Simons, Gary F., William D. Lewis, Scott Farrar, D. Terence Langendoen, Brian Fitzsimons, and Hector Gonzalez. 2004b. "The Semantics of Markup: Mapping Legacy Markup Schemas to a Common Semantics." *Proceedings of the 4th Workshop on NLP and XML* (*NLPXML-2004*), 25–32. Association for Computational Linguistics.

Sperberg-McQueen, Michael, and Lou Burnard, eds. 1994. *Guidelines for Electronic Text Encoding and Interchange* (*TEI P3*). Chicago: The Association for Computers in the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association for Linguistic and Literary Computing (ALLC).

Xia, Fei, William D. Lewis, Michael W. Goodman, Joshua Crowgey, and Emily M. Bender. 2014. "Enriching ODIN." *Proceedings of the 9th International Conference on Language Resources and Evaluation,* 3151–3157. Reykjavik, Iceland.

Xia, Fei, William D. Lewis, Michael W. Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey, and Emily M. Bender. 2016. "Enriching a Massively Multilingual Database of Interlinear Glossed Text." *Language Resources and Evaluation* 50:321–349.

3

Management, Sustainability, and Interoperability of Linguistic Annotations

Nancy Ide

Introduction

In recent years, a noticeable upswing has occurred in linguistic annotation activity, which has expanded to cover a wide variety of linguistic phenomena. At the same time, the number and size of linguistically annotated language resources has increased dramatically, together with a proliferation of annotation tools to support the creation and storage of labeled data, various means for collaborative and distributed annotation efforts, and the introduction of crowdsourcing mechanisms, such as Amazon Mechanical Turk. All of this has created a need to manage and sustain these resources, as well as to find ways to enable them to be repeatedly reused and merged with other resources.

What Is Linguistic Annotation?

Linguistic annotation involves the association of descriptive or analytic notations with language data. The raw data may be textual, drawn from any source or genre, or it may be in the form of time functions (audio, video, and/or physiological recordings). The annotations themselves may include transcriptions of all sorts (ranging from phonetic features to discourse structures), part-of-speech and sense tags, syntactic analyses, "named entity" labels, semantic role labels, time and event identification, co-reference chains, discourse-level analyses, and many others. Resources vary in the range of annotation types they contain: Some resources contain only one or two types, while others contain multiple annotation "layers," or "tiers," of linguistic descriptions.

Linguistic annotation of language data was originally performed in order to provide information for the development and testing of linguistic theories, or, as it is known today, corpus linguistics. At the time, considerable time and effort was required to annotate data with even the simplest linguistic phenomena, and the annotated corpora available for study were quite small. Over the past three decades, however, advances in computing power and storage, together with development of robust methods for automatic annotation, have made linguistically annotated data increasingly available in ever-growing quantities.

As a result, these resources now serve not only linguistic studies but also the field of natural language processing (NLP), which relies on linguistically annotated text and speech corpora to evaluate new human language technologies and, crucially, to develop reliable statistical models for training these technologies.

A linguistic annotation scheme is composed of two main parts: the scheme's *semantics*, which specify the categories and features that label and provide descriptive information about the data with which they are associated, and the scheme's *representation*, which is the physical format in which the annotation information is represented for consumption by software (and, in some cases, by humans as well). Historically, designers of linguistic annotation schemes have focused on determining the appropriate categories and features to describe the phenomenon in question and have paid less attention to the eventual physical representation of the annotation information, with possibly unintended results when constraints imposed by the physical representation affect choices for the conceptual content of an annotation scheme. In recent years, the need to compare and combine annotations, as well as to use them in software environments for which they may have not been originally designed, has increased, leading to the awareness that a conceptual scheme may be represented in any of a variety of different physical formats and/or transduced from one to the other.

Both the syntax and the semantics of an annotation scheme involve choices that are, to some extent, arbitrary, but that nevertheless have ramifications for their usability. With regard to the physical format, the most significant choice is whether to insert the annotation information into the data itself or to represent it in *standoff* form—that is, provided in a separate document with links to the positions in the original data to which each annotation applies.

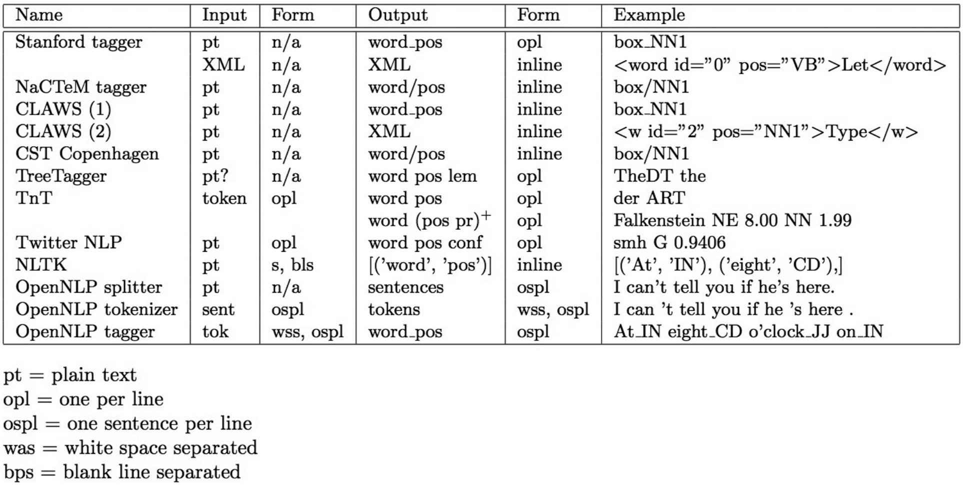
History

In the mid-twentieth century, linguistics was practiced primarily as a descriptive field, studying structural properties within a language and typological variations between languages. This work resulted in fairly sophisticated models of the various informational components comprising linguistic utterances. As in the other social sciences, the collection and analysis of data were also subjected to quantitative techniques from statistics, and in the 1940s, linguists such as Leonard Bloomfield and others were starting to think that language could be explained in probabilistic and behaviorist terms. At the same time, in the related and emerging field of NLP Warren Weaver suggested using computers to translate documents between natural human languages; in 1949 he produced a memorandum entitled "Translation" (Weaver 1955), which outlined a series of methods for that task. Empirical and statistical methods remained popular throughout the 1950s, and Claude Shannon's information-theoretic view to language analysis provided a solid quantitative approach for modeling qualitative descriptions of language. However, datasets were generally so small that it was not possible to extract statistically significant patterns to support probabilistic approaches, and as a result, linguistically annotated corpora did not play a major role in the first years of NLP (1950s–1960s).

During the 1960s, there was a general shift in the social sciences, particularly in the United States, from data-oriented descriptions of human behavior to introspective modeling of cognitive functions. As part of this new attitude toward human activity, the US linguist Noam Chomsky focused on both a formal methodology and a theory of linguistics that not only ignored quantitative language data, but also claimed that it was actually misleading for formulating models of language behavior. Chomsky's view was influential in the United States throughout the next two decades, largely because the formal approach enabled the development of extremely sophisticated rule-based language models using mostly introspective (or self-generated) data, thus providing an attractive alternative to creating statistical language models on the basis of relatively small datasets of linguistic utterances from the existing corpora in the field. In NLP, the flourishing field of artificial intelligence (AI) began to attack the problem of language understanding and, in the spirit of the times, AI proponents abandoned empirical methods and grounded their design of language processing systems in formal theories of human language understanding, which in turn they attempted to model. IBM's championing of statistical methods for speech processing in the 1970s and '80s was one of the few efforts that bucked this trend during that era. Reasonably large linguistically annotated resources were relatively rare; a well-known exception is the one-million-word Brown Corpus of Standard American English (Kuĉera and Francis 1967). In the 1970s, the Brown Corpus was the object of what is arguably the first modern linguistic annotation project, which added part-of-speech annotations.[1] Like the Brown Corpus, corpora developed in the 1970s and '80s were typically annotated only for part-of-speech, because the lack of reasonably accurate automatic methods as well as the high cost of manual annotation did not permit the production of sufficiently large corpora containing annotations for other linguistic phenomena, such as syntax.[2]

All this changed in the mid- to late-1980s, when large-scale language data resources started to become available. This led to a proliferation of linguistic annotation projects, most of them still focused on part-of-speech (or richer morphosyntactic) annotations, and in turn this spearheaded the reintroduction of probabilistic methods for automatic annotation based on statistical data derived from the corpus. The first major effort of this kind produced morphosyntactic and syntactic annotations of the one-million-word Lancaster-Oslo-Bergen (LOB) corpus of English (Garside 1987). Building on this work, the Penn Treebank project (Marcus, Marcinkiewicz, and Santorini 1993) produced a one-million-word corpus of *Wall Street Journal* articles annotated for part-of-speech and skeletal syntactic annotations and, later, also annotated for basic functional information (Marcus et al. 1994). Automatically produced annotations subsequently validated by humans (in whole or in part) were used to create several other major corpora in the 1990s, including the 100-million-word British National Corpus (Clear 1993), released in 1994; corpora produced by the MULTEXT project (1993–96; Ide and Véronis 1994) and its follow-on, MULTEXT-EAST (1994–1997; Erjaveç and Ide 1998), which provided parallel aligned corpora in a dozen Western and Eastern languages annotated for part-of-speech; and the PAROLE and SIMPLE corpora,[3] which included part-of-speech tagged data in fourteen European languages. Following these efforts, syntactic treebanks for a wide variety of languages (e.g., Swedish, Czech, Chinese, French, German, Spanish, Turkish, Italian) proliferated over the next decade, together with corpora annotated for other phenomena, such as word sense annotations (SemCor; Landes, Leacock, and Tengi 1998), which similarly engendered the development of comparably annotated corpora in other languages (Bentivogli, Forner, and Pianta 2004; Lupu, Trandabăţ, and Husarciuc 2005; Bond et al. 2012).

During this period, linguistic annotation was often motivated by the desire to study a given linguistic phenomenon in large bodies of data, so annotation schemes typically reflected a specific linguistic theory directly. Designers of linguistic annotation schemes focused on determining the appropriate categories and features to describe the phenomenon in question and paid less attention to the eventual *physical representation* for the annotations in the resource. Insofar as physical format was considered, the chief criterion for determining them was invariably the ease of processing by software that would use the output. For example, early formats for phenomena such as part of speech often output one word per line, separated from its part of speech (POS) tag by a special character such as an underscore or a slash (DeRose 1988; Church 1988; see figure 3.1). Syntactic parsers that produced constituency analyses typically used what has come to be known as the



Name	Input	Form	Output	Form	Example
Stanford tagger	pt	n/a	word_pos	opl	box_NN1
	XML	n/a	XML	inline	<word id=”0” pos=”VB”>Let</word>
NaCTeM tagger	pt	n/a	word/pos	inline	box/NN1
CLAWS (1)	pt	n/a	word_pos	inline	box_NN1
CLAWS (2)	pt	n/a	XML	inline	<w id=”2” pos=”NN1”>Type</w>
CST Copenhagen	pt	n/a	word/pos	inline	box/NN1
TreeTagger	pt?	n/a	word pos lem	opl	TheDT the
TnT	token	opl	word pos	opl	der ART
			word (pos pr)$^+$	opl	Falkenstein NE 8.00 NN 1.99
Twitter NLP	pt	opl	word pos conf	opl	smh G 0.9406
NLTK	pt	s, bls	[(’word’, ’pos’)]	inline	[(’At’, ’IN’), (’eight’, ’CD’),]
OpenNLP splitter	pt	n/a	sentences	ospl	I can’t tell you if he’s here.
OpenNLP tokenizer	sent	ospl	tokens	wss, ospl	I can ’t tell you if he ’s here .
OpenNLP tagger	tok	wss, ospl	word_pos	ospl	At_IN eight_CD o’clock_JJ on_IN

pt = plain text
opl = one per line
ospl = one sentence per line
was = white space separated
bps = blank line separated

Figure 3.1
Different formats for part-of-speech annotation produced by several tools.

"Penn Treebank format," which brackets and nests constituents with parentheses, LISP-style (Marcus, Marcinkiewicz, and Santorini 1993; Charniak 2000; Collins 2003).

Dependency parsers often used a line-based format that provides the syntactic function and its arguments in specified fields. Interestingly, these early formats for POS tagger and parser output have remained in use, with very little variation, up to the present day, primarily in the output of POS taggers; see, for example, the Stanford taggers and parsers for multiple languages,[4] TreeTagger,[5] and TnT.[6] Such formats rely heavily on white space and line breaks, together with occasional special characters, to delineate elements of the analysis (e.g., individual tokens and part-of-speech tags). As a result, software intended to use these formats as input must be programmed to understand both the meaning of these separators and the nature of the information in each field.

The separation between conceptual content and physical representation has not always been taken into account when schemes are designed, with possibly unintended results; for example, a representation format may impose limits on the complexity of the information that can be included, or can even force the conflation of information into cryptic labels that may be impossible to later disentangle. In recent years, the need to compare and combine annotations, as well as to use them in software environments for which they may have not been originally designed, has increased, leading to the awareness that a conceptual scheme may be represented in any of a variety of different physical formats and/or transduced from one to the other. Experience with annotated data that is difficult to transduce or modify has engendered annotation "best practices" that dictate that annotation information be both explicit (so that it can be readily retrieved) and flexible (so that other information can be substituted or added).

As the need for reliable automatic annotation for larger and larger bodies of data increased, there sometimes arose a tension between the requirements for accurate automatic annotation and a comprehensive linguistic accounting that could contribute to validation and refinement of the underlying theory. An early example in the 1990s is the Penn Treebank project's reduction and modification of the part-of-speech tagset developed for the Brown Corpus, in order to obtain more accurate results from automatic taggers and parsers. In the following decades, machine learning arose as the central methodology for NLP; therefore, some annotation projects began to design schemes incrementally, relying on iterative training and retraining of learning algorithms to develop annotation categories and features in order to best tune the scheme to the learning task (see, for example, Pustejovsky and Stubbs 2012)—in a sense shifting 180 degrees from a priori scheme design based on theory to a posteriori scheme development based on data and potentially limited by constraints on feature identification. Despite the increasing prevalence of this approach, there has been little discussion of the impact and value of iterative scheme development in the service of machine learning.

The Rise of Standards

Over the past 30 years, generalized solutions for representing annotated language data—that is, solutions that can apply to a wide range of annotation types and therefore allow for combining multiple layers and types of linguistic information—have been proposed.[7] The earliest format of note is the Standard Generalized Markup Language (SGML; ISO 8879:1986), which was introduced in 1986 to enable sharing of machine-readable documents, with no special emphasis on (or even concern for) linguistically annotated data. Like its successor the Extensible Markup Language, or XML (Bray et al. 2006), SGML defined a "meta-format" for marking up (meaning annotating) electronic documents consisting of rules for separating markup (tags) from data (by enclosing identifying names in angle brackets) and also for providing additional information in the form of attributes (features) on those tags.[8] SGML also specified a context-free language for defining tags and the valid structural relations among them (nesting, order, repetition, etc.) in an *SGML Document Type Definition* (DTD) that is used by SGML-aware software to validate the appropriate use of tags in a conforming document. XML replaced the DTD with the XML schema, which performs the same function as well as some others.

The Text Encoding Initiative (TEI)[9] Guidelines, first published in 1992, defined a broad range of SGML (and, later, XML) tags along with accompanying DTDs for encoding language data. However, the TEI was from its beginnings intended primarily for humanities data and does not provide guidelines for representing many phenomena of interest for linguistic annotation. Therefore, in the mid-1990s, the EU EAGLES project[10] defined the Corpus Encoding Standard (CES; Ide 1998), a customized application of the TEI providing a suite of SGML DTDs for encoding linguistic data and annotations, which was later instantiated in XML (XCES; Ide, Bonhomme, and Romary 2000). In part as a result, SGML (and, later, XML) began appearing in annotated language data during the mid-1990s—for example, in corpora developed in European Union–funded projects such as PAROLE, data used in the US-DARPA Message Understanding Conferences (MUC; Grishman and Sundheim 1995), and the TIPSTER annotation architecture (Grishman 1998) defined for the NIST Text Retrieval Conferences (TREC),[11] which included a CES-based SGML format for exporting output from information extraction tasks. SGML and XML were also adopted by major annotation frameworks developed during this period, such as GATE[12] and NITE,[13] for import and export of data.

Although widely adopted, XML as an in-line format for representing linguistic annotations did not solve the reusability problem, for several reasons. First and foremost, XML requires that in-line tags are structured as a well-formed tree, thus disallowing annotations that form overlapping hierarchies and making cumbersome connections between discontiguous portions of the data. In addition, like all in-line formats, the insertion of annotation information directly into the data imposes linguistic interpretations that may not be desired by other users. This includes segmental information—for instance, delineation of token boundaries in-line, whether by surrounding a string of characters with XML tags or by separating it with white space, line breaks, or other special characters—as well as the inclusion of specific annotation labels and features. To solve this problem, in 1994 the notion of *stand-off annotation* was introduced in the CES,[14] wherein annotations are maintained in separate documents and linked to appropriate regions of primary data, rather than interspersed in the primary data or otherwise modifying them to reflect the results of processing. This allows various annotations for the same phenomenon to coexist, including variant segmentations (e.g., tokenizations), as well as alternative analyses produced by different processors and/or using different annotation labels and features.

Annotation Graphs (AG; Bird and Liberman 2001), introduced in 2001, are a standoff format that represents annotations as labels on edges of multiple independent graphs defined over text regions in a document. Because the model was developed primarily with speech data in mind, the regions are typically defined between points on a timeline, although this is not necessary. However, because each annotation type or layer is represented by using a separate graph, the AG format is not well-suited to representing hierarchically based phenomena such as syntactic constituency.[15]

Over the past decade, there has been an increasing convergence of practice for representing linguistic annotations in the field, with the aims of ensuring maximal reusability and also reflecting advances in our understanding of possible means to best structure and organize data, especially Linked Data intended for access and query over the web. In addition to the use of standoff rather than in-line annotations, the focus has shifted from identifying a single, universal format to defining an underlying data model for annotations that can enable trivial, one-to-one mappings among representation formats without loss of information. The most generalized implementation of this approach is the International Standards Organization (ISO) 24612 Linguistic Annotation Framework (LAF; ISO 24612:2012; Ide and Suderman 2014), which was developed over the past fifteen years to provide a comprehensive and general model for representing linguistic annotations. To accomplish this, LAF was designed to capture the general principles and practices of both existing and foreseen linguistic annotations, including annotations of all media types such as text, audio, video, image, and so on, in order to allow for variation in annotation schemes, while at the same time enabling comparison and evaluation, merging of different annotations, and development of common tools for creating and using annotated data.

LAF specifies a set of fundamental architectural principles, including the clear separation of primary data from annotations (i.e., standoff annotation); separation of annotation structure (i.e., physical format) and of annotation content (the categories or labels used in an annotation scheme to describe linguistic phenomena); and a requirement that all annotation information be explicitly represented rather than building knowledge about the function of separators, position, and the like into processing software. LAF also defined an abstract data model for annotations, consisting of an acyclic digraph decorated with feature structures, grounded in *n*-dimensional regions of primary data.

The LAF data model and architectural principles, which in large part simply brought together existing best practices from a variety of sources, significantly influenced subsequent development of models and strategies to render linguistic annotations maximally interoperable. As a result, most general-purpose physical formats developed over the past decade embody virtually all of LAF's principles. Formats to enable interoperability within large systems and frameworks have also followed many of the same principles and practices—for example, the Unstructured Information Management Architecture's (UIMA; Ferrucci and Lally 2004) Common Analysis System (CAS), and the recently developed Language Applications Grid Interchange Format (LIF; Verhagen et al. 2015), which is a JSON-LD-based format designed for interchange among language processing web services (JSON: JavaScript Object Notation). The convergence of practice around the graph-based data model has led to the realization of increased compatibility of formats via mapping, and, as a result, transducers among formats are increasingly available that allow for the processing of annotated language resources by different tools and for different purposes (e.g., ANC2Go [Ide, Suderman, and Simms 2010], Pepper [Zipser and Romary 2010], and transducers available with DKPro[16] and the Language Applications [LAPPS] Grid.).[17]

However, one widely used format that was developed over the past two decades does not follow LAF's principles. The desire for processing ease and readability fostered development of a simple, column-based format for annotations for use in the Conference on Natural Language Learning (CoNLL) exercises. Most recently, a major project has developed a standard based on this format called CoNLL-U, the Universal Dependencies (UD) annotation format (Nivre et al. 2016). In this scheme, annotations are rooted in a fixed tokenization, are itemized in a single column, and are not linked to primary data. Each column corresponds to a defined annotation type, indicating whether the token in each row "begins" the annotation, is "inside" it, or is "outside" it.[18] Nested annotations, such as a constituency parse, are difficult to represent in this format without exploding the number of columns; to be fair, UD is intended primarily for dependency parses that do not present this problem. Alternative annotations of a given type cannot be represented easily, because each column in the UD format has predefined content, and because each row provides information for the token at its head. Other kinds of representation require even more gymnastics, if possible at all: for example, linking a given token such as the German "im" to its full form "in dem," which should be represented in two separate lines, thus disturbing the one-item-per-numbered-row scheme. Mapping UD or any similar column-based format to almost any other format is problematic at best, thus hampering interoperability. However, the ease of processing and the readability of this format have made these formats highly popular, and they are not likely to be abandoned anytime soon.

Interoperability as a Focus

Over the past fifteen years, what was referred to as “reusability” in the late 1990s came to be known as “interoperability.” During this period, the need for interoperability for linguistically annotated resources became increasingly urgent, as more and more language data were being annotated for more than one type of linguistic phenomenon, and as the need to use these annotations together was becoming more apparent. An experiment in the mid-2000s served to bring the need for annotation interoperability to the fore, especially in the United States, where it had been less a concern than in Europe: A project funded by the US National Security Agency called for annotation projects at labs around the states to annotate the same data (the 10,000-word Language Understanding [LU] corpus, or “Boyan 10K”) for a wide variety of linguistic phenomena in order to study inter-level interactions. The annotations included syntax, semantic roles, opinion, committed belief, and others. Ultimately, experts determined that it was impossible to combine the annotations, because of differences in formats, labels for the same phenomena, conceptions of what is a relation and what is an object, and a loss of information implicit in the original representations when combining was attempted. The most insurmountable problem was a huge variation in tokenization practices, which are often minimally documented, if at all.

Beyond these difficulties, the definition of what it means for linguistic annotations to be interoperable is unclear, but a clear definition is obviously necessary in order both to assess the current state of interoperability in the field and to measure our progress toward achieving interoperability in the future. What is needed, then, is an *operational definition*, which identifies one or more specific observable conditions or events that can be reliably measured, and tells where the results of the process are replicable.

Broadly speaking, interoperability can be defined as a measure of the degree to which diverse systems, organizations, and/or individuals are able to work together to achieve a common goal. For computer systems, interoperability is typically defined in terms of *syntactic interoperability* and *semantic interoperability*. Syntactic interoperability relies on specified data formats, communication protocols, and the like to ensure communication and data exchange. The systems involved can process the exchanged information, but there is no guarantee that the interpretation is the same. Semantic interoperability, by contrast, exists when two systems have the ability to automatically interpret exchanged information meaningfully and accurately and can produce useful results via deference to a reference model of common information exchange. The content of the information exchange requests is unambiguously defined: What is sent is the same as what is understood. More formally, semantic interoperability of data categories C_1 and C_2 is the capability of two annotation consumers to interchange annotation a_1 using C_1 and annotation a_2 using C_2 via a function f that maps C_1 to C_2, such that an analysis of C_2 is identical to the analysis of $f(C_1)$; that is, an analysis should produce the same result for two different but interoperable data categories.

For language resources, the focus today is increasingly on semantic rather than syntactic interoperability. That is, the critical factor is seen to be the accurate and consistent interpretation of exchanged data, rather than the ability to process the data immediately without modifying their physical format. The reasons for this are several, but first and foremost is the existence of large amounts of legacy data in several syntactic formats, coupled with the continued production of resources representing linguistic information in varied, but mappable, ways. Indeed, to ensure interoperability for language resources, the trend in the field is to specify an *abstract data model* for structuring linguistic data to which syntactic realizations can be mapped, together with a mapping to a set of *linguistic data categories* that communicate the information (linguistic) content. In the context of language resources, then, we can define syntactic interoperability as the ability of different systems to process (read) exchanged data either directly or via trivial conversion. Semantic interoperability for language resources is virtually the same as for software systems: It can be defined as the ability of systems to interpret exchanged linguistic information in meaningful and consistent ways, by reference to a common set of categories.

Semantic interoperability for linguistic annotation has proven to be more elusive than syntactic interoperability. As early as the 1990s, efforts were devoted to establishing standard sets of data categories, most notably within the European EAGLES/ISLE project,[19] which developed standards for morphosyntax, syntax, subcategorization, text typologies, and others. However, none of these standards has achieved universal acceptance and use. Recent large-scale efforts addressing standardization of data categories include those within ISO/TC 37/SC4 (Language Resource Management), which in 2004 proposed a registry accommodating the needs of linguistic annotation (Ide and Romary 2004) and subsequently implemented ISOcat (Kemps-Snijders et al. 2009), an online repository that is accessible and extensible with new data categories by the community. Recently, the ISOcat categories relevant for linguistic annotation were migrated to the *CLARIN Data Concept Registry.*[20] Other efforts include OLiA (Chiarcos 2012), a repository of annotation terminology for various linguistic phenomena intended to apply across multiple languages, and the Web Service Exchange Vocabulary (Ide et al. 2014b) under development within the Language Applications (LAPPS) Grid project (Ide et al. 2014a).

Despite these repeated efforts, at the present time no universally accepted set of categories exists, nor does even agreement on what the categories should be. However, some consensus has been reached, at least among schemes intentionally tailored to meet the needs of common NLP tools, which rely on some relatively common practices that have evolved over the years. These commonalities typically refer to attribute types, such as "part-of-speech," "constituent," "semantic role," and "relation" and leave open the range of valid values. This avoids some of the nastier kinds of mapping problems by pushing off problems of harmonization among specific values to another phase or mechanism; for example, tools may be required to provide metadata about the itemized tagsets they input and/or output (e.g., the Penn Treebank part of speech tags or the PropBank scheme of semantic role assignment) that can be checked for consistency at runtime. Other types of annotation have a fairly consistent (or at least easily mappable) set of categories, such as nounchunk and verbchunk, coreference (mentions, representative), common subsets of named entities (person, organization, location, date), dependencies (head, dependent), and so forth. Still, full consensus on linguistic categories and values is unlikely to be achieved anytime soon, if at all. As with syntactic interoperability, the best path may be to find means to allow flexibility while maintaining the ability to map among categories.

Conclusion

At this time, there is convergence within the community of various means to achieve annotation interoperability and a general willingness to pursue and ensure such means. However, it is difficult to identify an obvious solution or even a clear path to follow in order to fully achieve it. New technologies will likely emerge that may affect the way we approach the interoperability problem, much as the development of the Semantic Web and its supporting RDF/OWL format have impacted data models for annotations over the past fifteen years. In the meantime, the plodding progress in pursuit of interoperability that has been made over the past three decades will continue, inching toward a solution that is as yet only distantly visible.

Notes

1. It is interesting to note that the Brown Corpus annotation project fostered the development of increasingly accurate automatic methods for part-of-speech tagging in order to avoid the painstaking work of manual validation.

2. The earliest automatic part-of-speech taggers include Greene and Rubin's TAGGIT (Greene and Rubin 1971), Garside's CLAWS (Garside 1987), DeRose's VOLSUNGA (DeRose 1988), and Church's PARTS (Church 1988).

3. http://nlp.shef.ac.uk/parole/parole.html.

4. http://nlp.stanford.edu/software/tagger.shtml.

5. http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/.

6. http://www.coli.uni-saarland.de/~thorsten/tnt/.

7. Several initiatives have focused on reusability of language data from the late 1980s onward.

8. Note that the Hypertext Markup Language (HTML) is an *application* of SGML/XML, in that it uses the SGML/XML meta-format to define specific tag names and document structure for use in creating web pages.

9. www.tei-c.org/.

10. http://www.ilc.cnr.it/EAGLES/browse.html.

11. http://www-nlpir.nist.gov/related_projects/tipster/trec.htm.

12. http://gate.ac.uk.

13. http://groups.inf.ed.ac.uk/nxt/index.shtml.

14. Originally called “remote markup”–see http://www.cs.vassar.edu/CES/CES1-5.html#ToCOview.

15. An ad hoc mechanism to connect annotations on different graphs was later introduced into the AG model to accommodate hierarchical relations.

16. http://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/.

17. http://lappsgrid.org.

18. The three possibilities are designated with “B,” “I,” and “O,” respectively; the CoNLL format is often called the “BIO” format as a result.

19. http://www.ilc.cnr.it/EAGLES96/browse.html.

20. https://openskos.meertens.knaw.nl/ccr/browser/.

References

Bentivogli, Luisa, Pamela Forner, and Emanuele Pianta. 2004. “Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus.” *Proceedings of the 20th International Conference on Computational Linguistics*, COLING ’04. Stroudsburg, PA: Association for Computational Linguistics.

Bird, Steven, and Mark Liberman. 2001. “A Formal Framework for Linguistic Annotation.” *Speech Communication* 33 (1–2): 23–60.

Bond, Francis, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. “Japanese SemCor: A Sense-Tagged Corpus of Japanese.” *6th International Conference of the Global WordNet Association (GWC-2012)*. Matsue.

Bray, Tim, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, Francois Yergeau, and John Cowan. 2006, September. “Extensible Markup Language (XML) 1.1 (Second Edition).” W3C Recommendation, W3C–World Wide Web Consortium.

Charniak, Eugene. 2000. “A Maximum-Entropy-Inspired Parser.” *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, 132–139. Stroudsburg, PA: Association for Computational Linguistics.

Chiarcos, Christian. 2012. “Ontologies of Linguistic Annotation: Survey and Perspectives.” *8th International Conference on Language Resources and Evaluation (LREC2012*, 303–310.

Church, Kenneth Ward. 1988. “A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.” *Proceedings of the Second Conference on Applied Natural Language Processing*, ANLC ’88, 136–143. Stroudsburg, PA: Association for Computational Linguistics.

Clear, Jeremy H. 1993. “The Digital Word.” In *The British National Corpus,* edited by George P. Landow and Paul Delany, 163–187. Cambridge, MA: MIT Press.

Collins, Michael. 2003. “Head-Driven Statistical Models for Natural Language Parsing.” *Computational Linguistics* 29 (4): 589–637 (December).

DeRose, Steven J. 1988. “Grammatical Category Disambiguation by Statistical Optimization.” *Computational Linguistics* 14 (1): 31–39 (January).

Erjaveç, Tomaž, and Nancy Ide. 1998. “The Multext-East Corpus.” *Proceedings of First International Conference on Language Resources and Evaluation*. 971–974.

Ferrucci, David, and Adam Lally. 2004. “UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment.” *Natural Language Engineering* 10 (3–4): 327–348.

Garside, Roger. 1987. “The CLAWS Word-Tagging System.” In *The Computational Analysis of English*, edited by Roger Garside, Geoffrey Leech, and Geoffrey Sampson, 30–41. London: Longman.

Greene, Barbara B., and Gerald M. Rubin. 1971. *Automatic Grammatical Tagging of English*. Department of Linguistics, Brown University.

Grishman, R. et al. 1998. “The Tipster Annotation Architecture.” *Proceedings of a Workshop Held at Baltimore, Maryland: October 13–15, 1998*, TIPSTER ’98. Stroudsburg, PA: Association for Computational Linguistics.

Grishman, Ralph, and Beth Sundheim. 1995. “Design of the MUC-6 Evaluation.” *Proceedings of the 6th Conference on Message Understanding*, MUC6 ’95, 1–11. Stroudsburg, PA: Association for Computational Linguistics.

Ide, Nancy. 1998. “Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora.” *Proceedings of the First International Language Resources and Evaluation Conference*, 463–70.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. “XCES: An XML-Based Encoding Standard for Linguistic Corpora.” *Proceedings of the Second International Language Resources and Evaluation Conference (LREC ’00).*

Ide, Nancy, James Pustejovsky, Christopher Cieri, Eric Nyberg, Di Wang, Keith Suderman, Marc Verhagen, et al. 2014a, May. “The Language Application Grid.” *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC ’14).* Reykjavik, Iceland: European Language Resources Association (ELRA).

Ide, Nancy, James Pustejovsky, Keith Suderman, and Marc Verhagen. 2014b. “The Language Application Grid Web Service Exchange Vocabulary.” *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT).* Dublin, Ireland.

Ide, Nancy, and Laurent Romary. 2004. “A Registry of Standard Data Categories for Linguistic Annotation.” *Proceedings of the Fourth International Language Resources and Evaluation Conference (LREC ’04)*, 135–138. Lisbon, Portugal.

Ide, Nancy, and Keith Suderman. 2014. “The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging.” *Language Resources and Evaluation* 48 (3): 395–418.

Ide, Nancy, Keith Suderman, and Brian Simms. 2010, May. “ANC2Go: A Web Application for Customized Corpus Creation.” *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC).* Paris: European Language Resources Association.

Ide, Nancy, and Jean Véronis. 1994. “MULTEXT: Multilingual Text Tools and Corpora.” *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, vol. I, 588–592. Kyoto, Japan.

ISO 8879:1986. Information Processing—Text and office systems—Standard Generalized Markup Language (SGML). Geneva: International Organization for Standardization.

ISO 24612:2012. Language Resource Management—Linguistic Annotation Framework. Geneva: International Organization for Standardization.

Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2009. “ISOCat: Remodelling Metadata for Language Resources.” *International Journal of Metadata, Semantics and Ontologies* 4 (November): 261–276.

Kuĉera, Henry, and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Landes, S., C. Leacock, and R. I. Tengi. 1998. "Building Semantic Concordances." In *WordNet: An Electronic Lexical Database*, edited by C. Fellbaum. Cambridge, MA: MIT Press.

Lupu, Monica, Diana Trandabăţ, and Maria Husarciuc. 2005, July. "A Romanian SemCor Aligned to the English and Italian MultiSemCor." *1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School, Proceedings*, 20–27. Cluj-Napoca, Romania.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, et al. 1994. "The Penn Treebank: Annotating Predicate Argument Structure." *Proceedings of the Workshop on Human Language Technology,* 114–119. Stroudsburg, PA: Association for Computational Linguistics.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics* 19 (2): 313–330 (June).

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, et al.. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection." *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 1659–1666. Paris: European Language Resources Association.

Pustejovsky, James, and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning—A Guide to Corpus-Building for Applications*. Sebastopol, CA: O'Reilly Media.

Verhagen, M., K. Suderman, D. Wang, N. Ide, C. Shi, J. Wright, and J. Pustejovsky. 2015. "The LAPPS Interchange Format." *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure WLSI (2015)*, 33–47. Kyoto, Japan: Springer International Publishing.

Weaver, Warren. 1949; rpt. 1955. "Translation." In *Machine Translation of Languages*, edited by William N. Locke and A. Donald Boothe, 15–23. Cambridge, MA: MIT Press. Reprinted from a memorandum Weaver wrote in 1949.

Zipser, Florian, and Laurent Romary. 2010. "A Model Oriented Approach to the Mapping of Annotation Formats Using Standards." *Workshop on Language Resource and Language Technology Standards, LREC 2010*. La Valette, Malta.

4

Linguistic Linked Open Data and Under-Resourced Languages: From Collection to Application

Steven Moran and Christian Chiarcos

In this chapter, we argue for the adoption and use of Linked Data for linguistic purposes and, in particular, for encoding, sharing, and disseminating under-resourced language data. We provide an overview of linguistic Linked Data in the context of creating datasets of under-resourced languages, and we describe what "under-resourced" language data are, focusing on lexical resources (wordlists and dictionaries) and annotated corpora (glosses and corpora). We discuss aspects of resource integration with two brief case studies of linguistic data sources that have been transformed into Linked Data. Lastly, we describe the state and the bandwidth of applications of Linked Open Data technologies to under-resourced languages in the general context of the Open Linguistics Working Group and the developing Linguistic Linked Open Data (LLOD) ecosystem.

Introduction

Language scientists are increasingly interested in and gleaning the benefits from integration and computing of under-resourced language data. Different users clearly have different data needs; for example, linguists working on typological theory may require broad but not necessarily deep datasets, while computational linguists typically require big data. Regardless, increased access to (interoperable) data is beneficial both for science and for enterprises; in the language resource community, it has been a subject of intense activity over the last three decades, marked by initiatives such as the TEI (since 1987),[1] ISO TC 37/SC 4 (since 2001),[2] the Open Linguistics Working Group (since 2010),[3] as well as several W3C Community and Business groups (the earliest being OntoLex,[4] since 2011).

A more recent trend in this field is the increased adoption of Linked Data for representing language resources, a technology that was originally designed to create synergies between data sources in the Web of Data. Linked Data has been the focus of several workshop series (e.g., Linked Data in Linguistics, annually since 2012; Multilingual Linked Open Data for Enterprises [MLODE], biannually since 2012). At the Ninth International Language Resource and Evaluation Conference (LREC-2014), Linked Data was announced as the hot topic in the language resource community, and, subsequently, it sparked increased activity in workshops, summer schools, and datathons, including the First Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL-2014, Reykjavik, Iceland, May 2014), the First Summer Datathon on Linguistic Linked Open Data (SD-LLOD 2015, Madrid, Spain, June 2015), the EUROLAN-2015 summer school on Linguistic Linked Open Data (Sibiu, Romania, July 2015), and the LSA Summer Institute workshop on the Development of Linguistic Linked Open Data (LLOD) Resources for Collaborative Data-Intensive Research in the Language Sciences (LLOD-LSA 2015, Chicago, July 2015).

Because the applications of Linked Data to language resources are manifold (Chiarcos, Nordhoff, and Hellmann 2012), an exhaustive and up-to-date survey is beyond scope for our contribution in this chapter. We thus take a particular focus on an original research problem in linguistics—that is, the investigation of under-resourced languages; we illustrate the potential of Linked Data for statistical approaches in typology and cross-linguistic multivariate methods for investigating worldwide linguistic and cultural diversity.

This involves dealing with the following questions:

- How can collaborative approaches and technologies be fruitfully applied to the development and sharing of resources for under-resourced languages?
- How can small language resources be reused efficiently and effectively, reach larger audiences, and be integrated into applications?
- How can these resources be stored, exposed, and accessed by end users and applications?
- How can research on under-resourced languages benefit from Semantic Web technologies, and specifically the Linked Data framework?

In this chapter, we argue for the benefits of creating and using Linked Data. In particular, Linked Data is a fruitful method for attaining interoperability and creating useful data disseminations of under-resourced languages. Many of these languages are spoken in areas only recently penetrated by technology such as cell phones, and this creates more data and therefore more economic opportunities for people using them.

First, we define what we mean by "under-resourced languages." Then we give a brief, nontechnical introduction to Linked Data and we home in on using Linked Data for linguistic purposes. Next, we provide two short case studies that illustrate the increased opportunity for collaboration when creating under-resourced language data and tools using Linked Data technologies. Later we describe a large in-progress collaborative dataset, the Linguistic Linked Open Data cloud (LLOD), and we introduce the Open Linguistics Working Group (OWLG), a movement led both by computer scientists and linguists aimed at increasing the synergy between research being done in small-scale circles (e.g., field workers and small-scale language documentation projects) and larger and often enterprise-driven initiatives like MLODE or LIDER[5] to support content analytics of unstructured multilingual data. We begin by describing why increased access to under-resourced languages is important. And we end with directions to additional information on Linguistic Linked Open Data, including some do-it-yourself guidelines.

What Are Under-Resourced Languages?

Linguistic Diversity

Even though our view is very far from complete, world-wide linguistic diversity is simply astounding (cf. Evans and Levinson 2009).[6] Given the state of the world's languages, many of which are either endangered or moribund,[7] it is a high priority to document and describe these languages.

With this picture in mind, another fact to bear in mind is the lack of data that would enable us to undertake broad quantitative studies on cross-linguistic diversity. Typologists have coped by using statistical sampling methods to infer characteristics from signals in the genealogical descent or areal contact between languages (Cysouw 2005). This lack of data on the world's languages is referred to as the bibliographic sampling bias. The World Atlas of Language Structures (WALS; Dryer and Haspelmath 2013) is a classic example, at least among typologists, of a convenience sample with over 150 variables, examples being "Word Order" and "Hand and Arm," that necessarily paints an incomplete picture of worldwide linguistic diversity, which in turn spurs qualitative or speculative explanations (McNew, Derungs, and Moran 2018).

The most detailed picture that exists regarding the linguistic documentation of the world's languages is the Glottolog (Nordhoff et al. 2013).[8] Glottolog contains a bibliography about what is currently known about the state of documentation of the world's languages and it is available as Linked Data (Hammarström et al. 2015).[9] But what is known about the documentation of the world's "under-resourced" languages, and how does Linked Data help us combine that data with already existing knowledge?

Under-Resourced Languages

It is clear that languages lacking any documentation whatsoever are "under-resourced," since they are simply *not resourced*, so to speak. There is, however, a notion that there is a set of languages somewhere between very minimally documented ones (say, one grammar or dictionary) and large well-documented languages (examples being Chinese, English, French, German, Russian, and Spanish). This set of languages has been given various labels in the literature. Perhaps the oldest is "low-density languages" (Jones and Havrilla 1998). The terms "medium-density" and "lower-density languages" have also been coined (e.g., Maxwell and Hughes 2006). The latter term specifically refers to "the amount of computational resources available, rather than the number of speakers any given language might have" (Maxwell and Hughes 2006; Meyers et al. 2007). The amount of accessible data, regardless of language-speaker quantities, is the theme that binds these various terms together.[10]

In the language resource community, various categories of "under-resourced" or "weakly supported" languages have been employed:

1. Lack of access to language data—a general lack of language documentation and description (no grammars, dictionaries, or corpora)
2. Lack of access to digital language data—resources exist but cannot easily be accessed
3. Lack of IT/NLP support
4. Limited interoperability of data and tools

For category 1, there are thousands of languages with minimal or no documentation at all. This fact is so clear that we need not list examples.[11]

Category 2 applies to languages for which materials exist but access to those materials is not possible. In the most basic case, there is a lack of access to a digital resource; for instance, some linguist created a corpus of language X using software Y that is now obsolete. Perhaps more often, the case of inaccessibility is due to other factors, such as unsupported character encodings, unavailable fonts, the lack of a standardized orthography, or simply inaccessible data (caused by copyright restrictions, because they are housed in private collections, or only a few paper copies exist, and so on). For audio and video data, the nontransformation from analog to digital (or future) formats, as happened with first reel-to-reel and then cassette tapes, hinders data access.

Category 3 of under-resourced language data is only relevant when the first two points have been addressed. Without localized digital data, language-specific IT/NLP applications cannot exist. In this regard, we see concretely where under-resourced languages lie, as for example the Hausa language which, with some 30 to 50 million speakers, does not possess the digital resources needed for doing basic Natural Language Processing (NLP) tasks.

Category 4 leads us to the final issue in defining under-resourced languages. Technologically, limited interoperability of data and tools is prevalent in many areas, such as tools and annotations, which use different formats and conventions. Until recently, the Russian language has been a prime example; despite being spoken by ~150 million people worldwide, it has until recently lacked large-scale corpora, annotation schemes, and experimental NLP tools . Since the publication of the syntactic annotations of the Russian National Corpus[12] in 2008, the situation is slowly improving. Yet, even the current lack of interoperable digital resources for developing NLP tools exemplifies the point about under-resourced languages raised by Maxwell and Hughes (2006): It is the lack of accessible digital data, not the population of speakers of a given language, that determines whether the language is under-resourced.

Linguistic Resources

Determining under-resourced languages from a computational perspective requires that the resources of a given language be quantified. In this regard, the METANET white papers (Rehm and Uszkoreit 2013) have summarized the status for (most) officially recognized languages in the European Union (EU). The picture is not particularly satisfying. Out of 30 languages, only English is classified as having good support in terms of language resources. In terms of language resources required by different subfields of NLP, half the EU languages have fragmentary support.[13] And only five EU national languages are said to have weak or no support in such resources.[14] Coverage is even more dismal within certain NLP subfields; for example, two-thirds of the languages have weak-to-no support for machine translation. Of course this is the NLP view, where the degree of resource support is estimated from experts' assessment of both the quality/size of digital text, speech, and parallel corpora and their annotations, and of the quality/coverage of machine-readable lexical resources and grammars.

Resource types adopted to define a language as being (under-)resourced in linguistics are somewhat different. Glottolog, as an example, reports on the known language documentation with a focus on grammars, grammar sketches, dictionaries, and wordlists. These resources usually come with qualitative analyses, that is, analyses written by linguists on the basis of certain theoretical preconceptions. By nature, the act of creating a description of a language imposes theoretical constraints on the material collected. In other words, no universally accepted theory exists for describing a language as a system or a model, hence these language resources, even when electronically available, are often not available in a machine-readable format and in any event are usually incompatible with each other. Similar interoperability issues exist between these resources and annotated corpora, with respect to machine-readable dictionaries and grammars required by the METANET definition of "weakly supported" languages.

However, several linguistic data structures have in fact been standardized, to various extents. We focus on lexical resources and annotated (corpus/gloss) data. The third major class of digital language resources—tools for automated and semiautomated annotation—is beyond the scope of this chapter, as it presupposes the availability of dictionaries or corpora.

Lexical Resources: Wordlists and Dictionaries

The wordlist is often considered the most basic linguistic data structure. This generalization is superficial and misses the fact that the wordlist may be more complex than a simple pair of words with labels, such as "gloss" and "word." Yet the question of what a gloss is, is important in defining the nature of the relationship between "gloss" and "word." Perhaps better defined in light of multilingual wordlists is the notion of a "concept" that maps to a particular language-specific form. For example, many languages collapse the notions of "hand" and "arm" (used by English speakers, for example) into one concept that is a single entity. Therefore, there is a mapping relation between certain concepts, as conceptualized in different languages, and their language-specific forms. The relationship between concept and form is neither a definition nor a translation, but rather what has been termed "counterpart" in multilingual comparative contexts (Good 2013).[15]

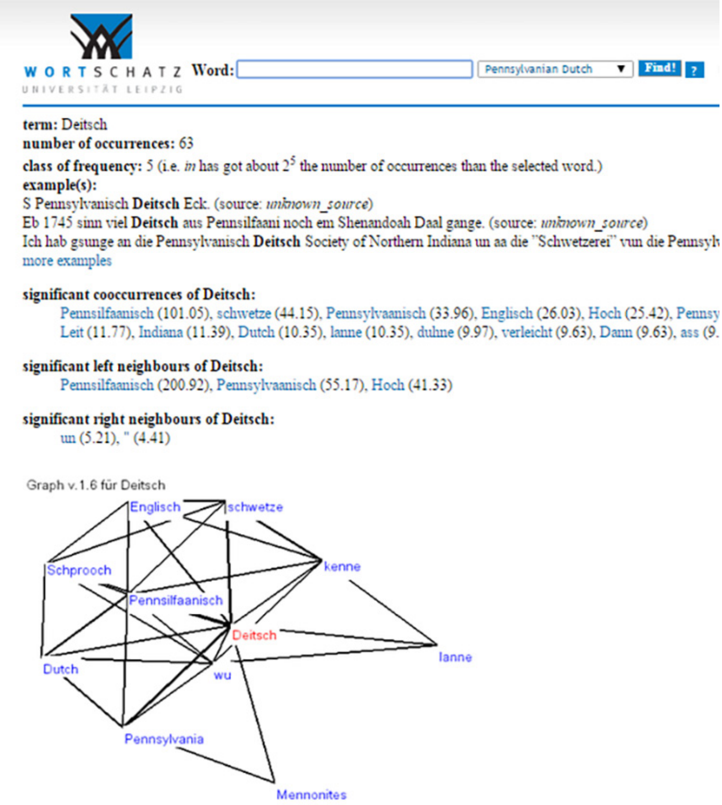
A dictionary is more detailed than a wordlist. It is typically idealized as a collection of form-to-meaning descriptions. Descriptions of forms are typically specified in culturally specific contexts (such as local flora and fauna), which makes it difficult to merge different dictionaries (or lexicons) into one large comparable multilingual source, like a multilanguage wordlist.

For languages that lack manually produced language resources but that come with considerable amounts of digitally available text, another type of lexical resource can be mentioned: frequency and collocation ("association") dictionaries that can be automatically derived from running text (Zock and Bilac 2004). One example is the Wortschatz portal,[16] which provides collocation and frequency dictionaries for 229 languages, including minor languages such as Manx (extinct), Neo-Aramaic (endangered), or Klingon (fictional). Figure 4.1 shows the example entry *Deitsch* "German" from Pennsylvania Dutch (a German dialect spoken in the United States) along with the information provided about it: frequency class (to estimate whether it is has grammatical or lexical function), examples, co-occurring words and frequent collocations, including words of the same semantic class (*Englisch*, *Dutch*, *Schprooch* "language"), related ethnic and geographic concepts (*Pennsylvania, Pennsilfaanisch, Mennonites*), and associated verbs (of speaking, *kenne* "to know," *lanne* "to learn," *schwetze* "to speak"). Although this information does not replace that in a traditional dictionary, it can be used as a tool to construct one, or to confirm the usage of an unknown word (Benson 1990). These resources are also useful for bootstrapping the development of multilingual lexical data translation graphs (cf. Kamholz, Pool, and Colowick 2014).

Annotated Data: Glosses and Corpora

In linguistically annotated data, examples are typically provided in the form of interlinear glossed text (IGT), a semi-standardized data structure comprising three or more lines that prototypically contain three items: an idiosyncratic transcription, a detailed linguistic interpretation (such as a morphological gloss or a part-of-speech tag), and a literal translation.[17] After identification (say, via regular expressions), IGT is automatically extracted from websites and online documents and then assigned an ISO 639-3:2007 language name identifier, derived from attributes identified in the source document. Searching across IGT of thousands of languages in varying detail is desirable, but since the transcription and annotation styles may differ from document to document, some additional layer of what may be called an ontological annotation is needed to logically and consistently define relations in the dataset (cf. Moran 2012a).

Taken a step further, the principle of glossing has been extended to the annotation of larger texts and even entire corpora, as for instance by using tools such as Toolbox.[18] By design, corpora are structured entities consisting of collections of primary data (texts, transcripts, image, audio, or video content), together with their metadata (author, source, date, location, language), and, usually, linguistic annotations as well. Modern corpora have been used as a tool for linguistic research since the Brown Corpus (Kučera and Francis 1967), which has since been compiled as a citation base for the *American Heritage Dictionary*, and which more recently became a cornerstone of corpus linguistics and NLP with the Penn Treebank (Taylor, Marcus, and Santorini 2003) and others.



term: Deitsch
number of occurrences: 63
class of frequency: 5 (i.e. *in* has got about 2^5 the number of occurrences than the selected word.)
example(s):
S Pennsylvanisch **Deitsch** Eck. (source: *unknown_source*)
Eb 1745 sinn viel **Deitsch** aus Pennsilfaani noch em Shenandoah Daal gange. (source: *unknown_source*)
Ich hab gsunge an die Pennsylvanisch **Deitsch** Society of Northern Indiana un aa die "Schwetzerei" vun die Pennsylv
more examples

significant cooccurrences of Deitsch:
Pennsilfaanisch (101.05), schwetze (44.15), Pennsylvaanisch (33.96), Englisch (26.03), Hoch (25.42), Pennsy
Leit (11.77), Indiana (11.39), Dutch (10.35), lanne (10.35), duhne (9.97), verleicht (9.63), Dann (9.63), ass (9.

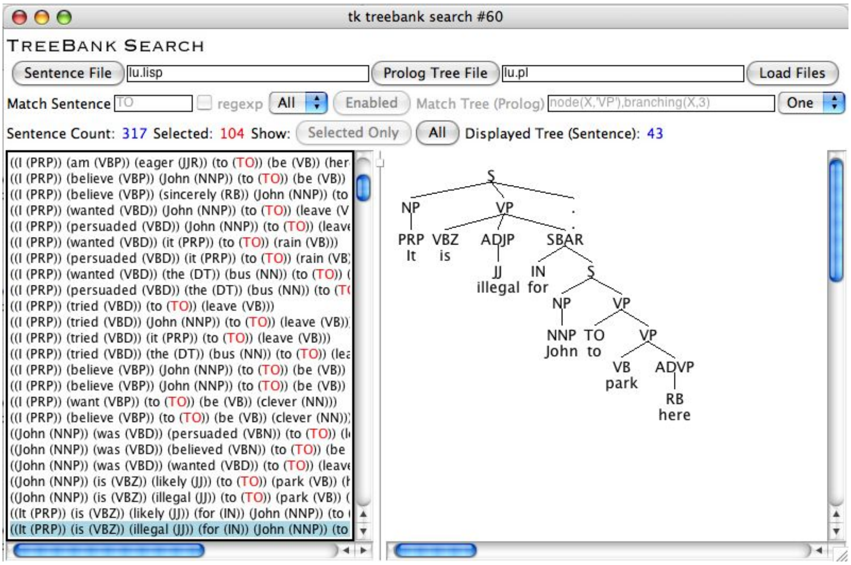
significant left neighbours of Deitsch:
Pennsilfaanisch (200.92), Pennsylvaanisch (55.17), Hoch (41.33)

significant right neighbours of Deitsch:
un (5.21), " (4.41)

Figure 4.1
Example word *Deitsch* (“German”) from Pennsylvania Dutch in the Wortschatz portal.

Taking the Penn Treebank as an example, typical annotations comprise lemmatization, morphosyntax (parts of speech, inflectional morphology), syntactic analyses (here phrase structure grammar, otherwise also nominal/clausal chunks or dependency analysis), and, for well-resourced languages, higher levels of analysis such as semantic roles (Kingsbury and Palmer 2002; Meyers, Reeves, Macleod, Szekely, et al. 2004), temporal relations (Pustejovsky et al. 2003), pragmatics (Carlson et al. 2002; Prasad et al. 2008), or co-reference (Pradhan et al. 2007)—in this case specialized subcorpora of the Penn Treebank. Figure 4.2 shows morphosyntactic and syntactic annotations of the Penn Treebank.[19]

For languages without annotated corpora, parallel corpora (such as the Bible, the Qur’an, various translated literature, technical or operational manuals, localization files from software distributions, or subtitles) can be used to bootstrap linguistic annotations via annotation projection (Yarowsky, Ngai, and Wicentowski 2001). Aligned syntactic annotations in a parallel corpus are shown in figure 4.3.[20]



Sentence File: lu.lisp
Prolog Tree File: lu.pl
Match Sentence: TO
Match Tree (Prolog): node(X,'VP'),branching(X,3)
Sentence Count: 317 Selected: 104 Displayed Tree (Sentence): 43

```
((I (PRP)) (am (VBP)) (eager (JJR)) (to (TO)) (be (VB)) (her
((I (PRP)) (believe (VBP)) (John (NNP)) (to (TO)) (be (VB))
((I (PRP)) (believe (VBP)) (sincerely (RB)) (John (NNP)) (to
((I (PRP)) (wanted (VBD)) (John (NNP)) (to (TO)) (leave (V
((I (PRP)) (persuaded (VBD)) (John (NNP)) (to (TO)) (leav
((I (PRP)) (wanted (VBD)) (it (PRP)) (to (TO)) (rain (VB)))
((I (PRP)) (persuaded (VBD)) (it (PRP)) (to (TO)) (rain (VB
((I (PRP)) (wanted (VBD)) (the (DT)) (bus (NN)) (to (TO)) (
((I (PRP)) (persuaded (VBD)) (the (DT)) (bus (NN)) (to (T
((I (PRP)) (tried (VBD)) (to (TO)) (leave (VB)))
((I (PRP)) (tried (VBD)) (John (NNP)) (to (TO)) (leave (VB))
((I (PRP)) (tried (VBD)) (it (PRP)) (to (TO)) (leave (VB)))
((I (PRP)) (tried (VBD)) (the (DT)) (bus (NN)) (to (TO)) (le
((I (PRP)) (believe (VBP)) (John (NNP)) (to (TO)) (be (VB))
((I (PRP)) (believe (VBP)) (John (NNP)) (to (TO)) (be (VB))
((I (PRP)) (want (VBP)) (to (TO)) (be (VB)) (clever (NN)))
((I (PRP)) (believe (VBP)) (to (TO)) (be (VB)) (clever (NN))
((John (NNP)) (was (VBD)) (persuaded (VBN)) (to (TO)) (l
((John (NNP)) (was (VBD)) (believed (VBN)) (to (TO)) (be
((John (NNP)) (was (VBD)) (wanted (VBD)) (to (TO)) (leav
((John (NNP)) (is (VBZ)) (likely (JJ)) (to (TO)) (park (VB)) (
((John (NNP)) (is (VBZ)) (illegal (JJ)) (to (TO)) (park (VB)) (
((It (PRP)) (is (VBZ)) (likely (JJ)) (for (IN)) (John (NNP)) (to
((It (PRP)) (is (VBZ)) (illegal (JJ)) (for (IN)) (John (NNP)) (to
```

Figure 4.2
Annotations of the Penn Treebank as visualized by TreeBank Search.

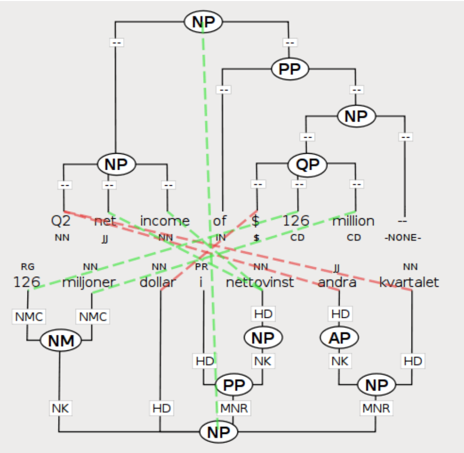


Figure 4.3
Parallel corpus with syntactic annotations and alignment as visualized by TreeAligner.

For languages with a great deal of digitally available text, but lacking NLP support, unsupervised NLP tools may be an option. These extend the concept of collocation extraction to unsupervised grammatical analysis (Clark 2003). However, as this information is only partially interpretable in terms of traditional grammatical categories, and requires considerable amounts of data, this is a current topic of research and beyond the scope of this chapter.

Summarizing, the structures of linguistic resources are manifold even within a single language, and for under-resourced languages resource development even requires links between such structured entities across different languages. Resource integration is thus not only a key problem for modern linguistics in general but also for under-resourced languages in particular.

Resource Integration

It is important to note that linguistic resources are *complex and structured* entities that are composed of different components that need to be integrated if interoperability is to be attained. For example, there is primary data (such as lexemes in a dictionary, text in a corpus, audio or video streams in multimedia corpora), secondary data (including natural language translations, such as glosses and their definitions in a dictionary, or the translation in a parallel corpus or a bilingual wordlist), grammatical analyses (such as in dictionaries, glosses, and annotations), and possibly cross-references (such as a keyword-in-context [KWIC] view in a corpus, a lookup facility from corpus to dictionary to compare the definition of a word, or a lookup facility from dictionary to corpus to provide real-world examples).

Out of this situation of inoperability of data sources and types emerges the challenge to represent (linguistic) data structures on a technical level. Varying solutions to the problem have been proposed, but they have often either been problem-specific (say, a domain-specific [lexicon] XML format via Toolbox) or what might be called “local” (that is, integration within a relational database, showing for instance how to store language and author-specific IGT examples). Each solution probably has its merits; the most widely known solutions have achieved a level of maturity or publicity that has led to their acceptance within their community.

Still, linguistic resources created in an idiosyncratic fashion are not easily reused, unless they can be (easily) integrated with other datasets. This is one of the core functionalities of Linked Data. But at the same time, Linked Data helps us to overcome the heterogeneity of existing formalisms for different local resources, such as dictionaries and corpora. However, existing infrastructures, resources, and tools will continue to be used, and it would be premature to suggest a general shift from existing technology to Linked Data. Instead, we delineate here ways that may be used to automatically convert an existing resource to Linked Data and demonstrate some of the benefits we have gleaned from this conversion.

To summarize, questions of how linguistic data types are transformed into Linked Data are as idiosyncratic as the projects or people who make the design decisions to convert from, say, a linguistic data type A to the Linked Data implementation B. We start with a brief overview of Linked Data and then we show how several datasets have been converted into Linked Data in the Linguistic Linked Open Data (LLOD) cloud.

Linked Data and Under-Resourced Language Data

Linked Data

Linked Data are a set of rules, or “best practices,” if you will, for publishing data on the web. Linked Data includes a set of protocols and standards, the purpose of which is to establish links between different datasets. Links are used here broadly; mechanisms provide ubiquitous URI resolution whether a user clicks on a link in his or her browser, or whether computer code automatically crawls through machine interpretable data.

The Linked Open Data paradigm postulates four rules for the publication and representation of web resources:

1. Referred entities should be designated by using URIs.
2. These URIs should be resolvable over HTTP.
3. Data should be represented by means of W3C standards (such as RDF; see below).
4. A resource should include links to other resources.

These rules facilitate information integration, and thus, interoperability, in that they require entities to be addressed in a globally unambiguous way (rule 1 above), that they can be accessed (rule 2) and interpreted (rule 3), and that entities that are associated on a conceptual level are also physically associated with each other (rule 4).

Linked Data is also focused on information integration, and in particular on structural and conceptual interoperability. Linked Data developers strive for **structural interoperability** to attain comparable formats and protocols to access both their own and others’ data. A goal is to use the same query language for different datasets, which the user can query across, with or without manipulating the underlying logic (or “semantics”) encoded into the (combined) dataset(s) (cf. Moran 2012b).

In the definition of Linked Data, the Resource Description Framework (RDF) receives special attention. RDF was designed to provide metadata about resources that are available either offline (as in books in a library) or online (e-books in a store). RDF provides a generic data model based on labeled directed graphs, which can be serialized in different formats. Information is expressed in terms of *triples*—consisting of a *predicate* (relation, i.e., a labeled edge) that connects a *subject* (i.e., a resource in the form of a labeled node) with its *object* (i.e., another resource or a literal or string). For example, the statement *Christian Chiarcos knows Steven Moran* might be (pseudo)-encoded as a single string consisting of the subject, predicate, and object triple:

```
Subject  http://www.acoli.informatik.uni-frankfurt.de/~chiarcos
Predicate  http://xmlns.com/foaf/0.1/knows
Object  http://www.comparativelinguistics.uzh.ch/de/moran.html
```

As shown, RDF resources (nodes)[21] are represented by *Uniform Resource Identifiers* (URIs), and they are therefore globally unambiguous in the Web of Data (as well as the

"Semantic Web"). Linked Data infrastructure allows resources hosted at different locations to refer to each other, which in turn creates a network of collections of data whose elements are densely interwoven.

Several linearizations for RDF data exist, which differ in readability and compactness. RDF/XML was the original standard for that purpose, but it has been largely replaced by Turtle, a more human-readable format. In Turtle, triples are written as sequences of subject, predicate, and object components, concluded with a final dot.

```
<http://www.acoli.informatik.uni-frankfurt.de/~chiarcos>
<http://xmlns.com/foaf/0.1/knows>
<http://www.comparativelinguistics.uzh.ch/de/moran>.
```

A more compact representation can be achieved using namespace prefixes instead of full URIs:

```
PREFIX acoli: <http://www.acoli.informatik.uni-frankfurt.de/~>
PREFIX cluzh: <http://www.comparativelinguistics.uzh.ch/de/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
acoli:chiarcos foaf:knows cluzh:moran .
```

Several database implementations for RDF data are available, and these can be accessed using **SPARQL** (Prud'hommeaux and Seaborne 2008), a standardized query language for RDF data. SPARQL uses a triple notation similar to Turtle, where properties and RDF resources can be replaced by variables. SPARQL was inspired by Structured Query Language (SQL), in which variables can be introduced in a separate SELECT block, and in which constraints on these variables are expressed in a WHERE block in a triple notation. Thus, for example, we can query for relations between two particular people:

```
SELECT ?relation
WHERE { acoli:chiarcos ?relation cluzh:moran . }
```

SPARQL does not only support running queries against individual RDF databases that are accessible over HTTP (so-called SPARQL endpoints), but it also allows users to combine information from multiple repositories (known as "federation"). RDF can thus be used both to *establish* a network (or cloud) of data collections, and to *query* that network directly.

In this way, Linked Data facilitates the resource accessibility and reusability on different levels (Ide and Pustejovksy 2010):

How to access (read) a resource? (Structural interoperability) Resources use comparable formalisms to represent and to access data (formats, protocols, query languages, etc.), so that they can be accessed in a uniform way and that their information can be integrated with each other.

How to interpret (understand) information from a resource? (Conceptual interoperability) Resources share a common vocabulary, so that linguistic information from one resource can be resolved against information from another resource, e.g., grammatical descriptions can be linked to a terminology repository.

How to integrate (merge) information from different resources? (Federation) Web resources are provided in a way that remote access is supported. Using structurally interoperable representations, a query language with federation support allows the user to run queries against multiple external resources within a single query, and thereby to integrate their information at query time.

In other words, structural interoperability means that resources can be accessed in a uniform way and that their information can be integrated with each other.

Conceptual interoperability is the goal to develop and (re-)use shared vocabularies for equivalent concepts. Shared vocabularies allow the user to run *the same query* across different datasets. Conceptual interoperability, also referred to as semantic interoperability, goes beyond using unified structural data formats and provides a type of label translation with an additional layer of Description Logics, as for example when using OWL-DL to encode datasets.[22]

Again, to make data structurally and conceptually interoperable (to varying degrees), the term *federation* refers to bringing structurally and conceptually interoperable datasets together on the web—publishing data already published on the web, preferably under an open license and with a query interface such as a SPARQL endpoint. Open data is part of the mission of the Open Linguistics Working Group (OWLG), which we describe later in this chapter. First, we highlight the data integration problem and then we discuss Linked Data in the contexts of under-resourced language data and NLP.

Under-Resourced Language Data

The tools used to produce language data and to create and disseminate detailed (and often computationally implemented)[23] linguistic analyses produce a rapidly increasing amount and depth of inoperable datasets. The breadth and depth of ongoing research projects range from many small-scale, single-scientist data collection projects (as in "linguist X works with the last remaining speaker of language Y") to smaller-to-medium-scale corpora collections (say, a one-million-word corpus of X), to larger-to-medium projects that combine many resources (such as CLLD),[24] to large-scale big-data producing efforts (Wiktionary, DBPedia, and the like).

Although the focus of each project differs, all of them gain from more or richer data sources. Among many, notable examples of collections that contain detailed data on under-resourced language data include the ANU Database (Donohue et al. 2013), AUTOTYP (Bickel and Nichols 2015), STEDT (Matisoff 2015), and PHOIBLE (Moran, McCloy, and Wright 2014). A tremendous amount of effort has been put into creating these rich datasets, which are often aimed at collecting linguistic diversity. Each dataset contains sets of languages that are under-resourced, but those data remain in project-specific formats, resulting in insufficient data access, possibilities for sharing, and integration for query and comparison.

Linked Data for Linguistics and NLP

For users wishing to create Linked Data for linguistics, we note that publishing Linked Data allows resources to be globally and uniquely identified such that they can be retrieved through standard web protocols. Moreover, resources can be easily linked to one another in a uniform fashion and thus become structurally interoperable. The five main benefits of Linked Data for linguistics and NLP can be stated as follows (Chiarcos et al. 2013):

Conceptual interoperability: Semantic Web technologies allow users to provide, to maintain, and to share centralized, but freely accessible terminology repositories. Reference to such terminology repositories facilitates conceptual interoperability, since different concepts used in the annotation are backed up by externally provided definitions; these common definitions may be employed for comparison or information integration across heterogeneous resources.

Linking through URIs: URIs provide globally unambiguous identifiers, and if resources are accessible over HTTP it is possible to create resolvable references to URIs. Different resources developed by independent research groups can be connected into a cloud of resources.

Information integration at query runtime (Federation): Along with HTTP-accessible repositories and resolvable URIs, it is possible to combine information from physically separated repositories in a single query at runtime; to wit, resources can be uniquely identified and easily referenced from any other resource on the web through URIs. Similar to hyperlinks in the HTML web, the so-called Web of Data created by these links allows for navigation along these connections, and thereby allows free integration of information from different resources in the cloud.

Dynamic import: When linguistic resources are interlinked by references to resolvable URIs instead of system-defined IDs (or static copies of parts from another resource), one should always provide access to the most recent version of a resource. For instance, for community-maintained terminology repositories like the ISO TC 37/SC 4 Data Category Registry (ISOcat; Windhouwer and Wright 2012; Wright 2004), new categories, definitions, or examples can be introduced occasionally, and this information is available immediately to anyone whose resources refer to ISOcat URIs. To preserve link consistency among Linguistic Linked Open Data (LLOD) resources, however, it is strongly advised to apply a proper versioning system such that backward-compatibility can be preserved: Adding concepts or examples is unproblematic, but when concepts are deleted, renamed, or redefined, a new version should be provided.

Ecosystem: RDF as a data exchange framework is maintained by an interdisciplinary, large, and active community, and it comes with a developed infrastructure that provides APIs, database implementations, technical support, and validators for various RDF-based languages, such as reasoners for OWL. For developers of linguistic resources, this ecosystem can provide technological support or off-the-shelf implementations for common problems; for example, a database can be developed to be capable of supporting flexible, graph-based data structures as necessary for multi-layer corpora (Ide and Suderman 2007).

To these, we may add that the distributed approach of the Linked Data paradigm facilitates the distributed development of a web of resources. It also provides a mechanism for collaboration between researchers who use data, employing shared sets of technologies. One consequence is the emergence of interdisciplinary efforts to create large and interconnected sets of resources in linguistics—and beyond.

These benefits are of particular importance to less-resourced languages. Through recent community efforts such as the OWLG and the emergence of the LLOD cloud, resources from many languages can now be:

- found through central metadata repositories (for the OWLG DataHub),
- accessed by traversing from one resource to another that is linked with it, and
- identified and documented through a set of shared vocabularies

It is important to note at this point that the mere availability of linguistic resources may already improve chances for not just finding but actually *developing* resources for additional under-resourced languages. For example, NLP tools, annotations, and machine-readable lexicons may be *ported* from one language to another, related one. This might not help language isolates, such as Basque or perhaps Etruscan, but it would greatly improve the situation of, say, Faroese if resources from Icelandic can be ported. A similar situation persists for the Bantu languages in Africa, for which a certain degree of NLP support has been achieved only in the nation of South Africa, whereas Bantu languages in most other countries further north have no support at all. In certain respects, these languages are relatively closely related, so that resource porting between languages may be an option.

Examples for such porting approaches include the analysis of Ugaritic (an ancient Semitic language spoken in the second millenium BCE) through resources originally developed for the morphological analysis of Hebrew (Snyder, Barzilay, and Knight 2010) or for approaches to performing character-based translation between related languages, as for example with orthography being “normalized” from a less-resourced language to another; the tool chain developed for the latter case can be applied to the former (Moran 2009; Tiedemann 2012). As a formalism to provide language resources in a structurally and conceptually interoperable way, Linked Data provides a potential cornerstone for future approaches on resource porting across varying languages and domains.

Case Studies

In defining under-resourced languages, we mentioned four key problems: (1) lack of access to language data, (2) lack of access to digital data, (3) lack of IT/NLP support, and (4) limited interoperability of data and tools. We can aim to increase the limited interoperability of data and tools by improving both the conceptual and structural interoperability of existing data sources. This can be undertaken with increased IT/NLP support between languages and projects, which can in turn be used to guide digitization efforts to (partially) compensate for the lack of lexical resources of under-resourced languages.

Efforts to improve conceptual and structural interoperability are exemplified by shared vocabularies; examples include Lexicon Model for Ontologies (*lemon*; McCrae et al. 2010; McCrae, Spohr, and Cimiano 2011; lexicons), Lexvo[25] (de Melo 2015) and Glottolog[26] (Hammarström et al. 2015; language identification), PHOIBLE Online[27] (Moran, McCloy, and Wright 2014; phonemes), and OLiA (Chiarcos 2008; annotations). Other efforts to increase the lack of lexical resources are exemplified by projects like QuantHistLing (see below), PanLex[28] (Kamholz, Pool, and Colowick 2014), and LiODi.[29] In this section we provide examples in the form of brief case studies.

QuantHistLing

Projects like QuantHistLing (Quantitative Historical Linguistics)[30] illustrate the effort needed to make linguistically diverse samples of lexical data available to a broad and computationally savvy audience. Any project must first identify the linguistic data sources (such as wordlists and dictionaries) that it wishes to use or to create. QuantHistLing has digitized about 200 source documents, most of them available only in print and many of them the sole resources available for the poorly described and under-resourced languages that they describe. Two examples, one of a comparative wordlist and the other of a bilingual dictionary, respectively, are shown in figure 4.4.

The digitization pipeline involves transforming printed sources into electronic sources (whether by OCR or by manual typing). Once sources exist in an electronic form, for dictionaries the interesting parts of each entry are identified, typically with source-specific regular expressions, to extract head words, translations, example sentences, and part-of-speech information. For wordlists, concepts and their glosses are extracted. Standoff annotations may be added to the data by project members; for example, the "dictinterpretation" data type is added by project members and may include manual corrections or other pertinent information.

The QuantHistLing project produces a simple data output format that contains metadata (prefixed with the symbol "@") and tab-delimited lexical output on a source-by-source basis.[31] An example is given in figure 4.5.

Using the comma-separated values (CSV) data as input, a simple script was written to transform the data into RDF. An RDF model that is specified in the Lexicon Model for Ontologies (*lemon*; McCrae et al. 2010; McCrae, Spohr, and Cimiano 2011) was created for the QuantHistLing data (Moran and Brümmer 2013). *Lemon* is an ontological model for modeling lexicons and machine-readable dictionaries for linking to both the Semantic Web and the Linked Data cloud. The QuantHistLing-lemon model is illustrated in figure 4.6.

Given the goals of QuantHistLing to uncover and clarify phylogenetic relationships between languages, the transformation of wordlist data and of dictionary data from numerous source documents to an RDF graph provides researchers with a structurally

Chocó

DR	hĩɾṹ
CT	hẽɾṹ
CM	híɾũ
TD	híɾã, ɓɨ́ɾɨ
EP	híɾũ
BA	bɨ́ɾɨ ekʰára
WM	bɨ

Chibcha

IK	kə́ttɨ
KO	kása
DM	kɨsá
CL	kássa
TN	kes-kára
BI	kixturə

Barbacoa

PA	tʃida
GU	katsik
TR	kaˈtsik
AW	mittɨ
TP	nede
CH	neepa

Kamsá

KS	ʃekuá-tçe

Quechua

IN	tʃáki

Arawak

WY	wóˀui (wa-óˀui)
AC	-íiba
CR	no-iipa
PP	wàabàli (wa-àbàli)
YC	weˀemá (wa-iˀimá)
TO	pititáβe, pititáwe†
CA	ħiipa
BN	-ipa
RE	-hiiˀpú

Tucano

TC	dɨˀpó-kã
WN	daˀˈpo-ro
PY	daˀˈpokã
WA	dɨˈpó
BR	dɨˈpo
TY	dɨˈpó
YR	ˈdɨ́po
DE	ˈgúbú-ru
SR	guˀˈbú
TA	rɨˈpó
CP	rɨˈpó
MA	gɨbo
BS	gɨbó
TM	ũˀˈpu-a
CU	kɨˈbó-ba
KG	ˈkũˀa-pɨ
SI	ˈgɨ̃õ-bɨ
SE	ˈkɨ̃õhawa
OR	ɨ̃õ-pɨ

Carib

CJ	ˈhuhu
YK	úʃi

Guahibo

PL	pe-táxu
GH	pe-táxu
CI	pe-táxu
JT	pe-tkút
GY	peh tɨ́ak

Sáliba-Piaroa

SL	ħaˀba

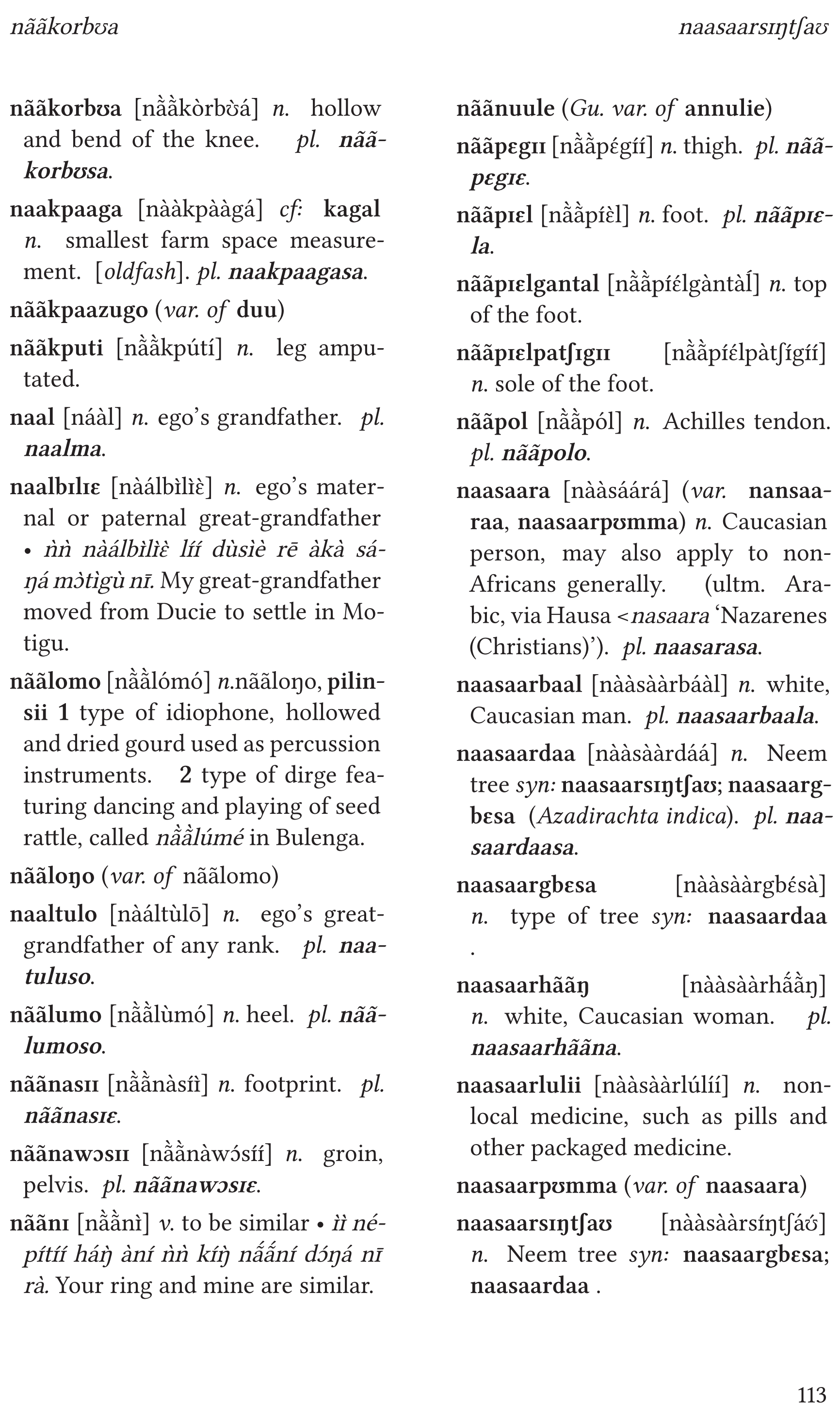
Macú-Puinave

PU	sim
NK	tʃɨ̃i^{4}at^{1}
KK	hit^{2}-tʃa^{4} daˀ^{4}
JU	tʃib

Witoto

MR	e.ɯ-dʒɯ
MN	é.ɯba
NP	e.ɯ́-ba
OC	ɯˀjóó(ga)
MU	tɨ́-ˀai
BO	(mé)-xtʰɯ́ˀaá
MÑ	tʰɯ́ˀaá, íhtʰʲɯ́ˀa

TO Giacone: ue-hépama



nã̀ã̀korbʊa *naasaarsɪŋtʃaʊ*

nããkorbʊa [nã̀ã̀kòrbʊ̀á] *n.* hollow and bend of the knee. *pl.* ***nããkorbʊsa***.

naakpaaga [nààkpààgá] *cf:* **kagal** *n.* smallest farm space measurement. [*oldfash*]. *pl.* ***naakpaagasa***.

nããkpaazugo (*var. of* **duu**)

nããkputi [nã̀ã̀kpútí] *n.* leg amputated.

naal [náàl] *n.* ego's grandfather. *pl.* ***naalma***.

naalbɪlɪɛ [nàálbɪ̀lɪ̀ɛ̀] *n.* ego's maternal or paternal great-grandfather • *ǹǹ nàálbɪ̀lɪ̀ɛ̀ líí dùsìè rē àkà sáŋá mɔ̀tìgù nī.* My great-grandfather moved from Ducie to settle in Motigu.

nããlomo [nã̀ã̀lómó] *n.*nããloŋo, **pilinsii** **1** type of idiophone, hollowed and dried gourd used as percussion instruments. **2** type of dirge featuring dancing and playing of seed rattle, called *nã̀ã̀lúmé* in Bulenga.

nããloŋo (*var. of* nããlomo)

naaltulo [nàáltùlō] *n.* ego's great-grandfather of any rank. *pl.* ***naatuluso***.

nããlumo [nã̀ã̀lùmó] *n.* heel. *pl.* ***nããlumoso***.

nããnasɪɪ [nã̀ã̀nàsɪ́ɪ̀] *n.* footprint. *pl.* ***nããnasɪɛ***.

nããnawɔsɪɪ [nã̀ã̀nàwɔ́sɪ́ɪ́] *n.* groin, pelvis. *pl.* ***nããnawɔsɪɛ***.

nããnɪ [nã̀ã̀nɪ̀] *v.* to be similar • *ɪ̀ɪ̀ népítíí háŋ̀ àní ǹǹ kíŋ̀ nã́ã́ní dɔ́ŋá nī rà.* Your ring and mine are similar.

nããnuule (*Gu. var. of* **annulie**)

nããpɛgɪɪ [nã̀ã̀pɛ́gɪ́ɪ́] *n.* thigh. *pl.* ***nããpɛgɪɛ***.

nããpɪɛl [nã̀ã̀pɪ́ɛ̀l] *n.* foot. *pl.* ***nããpɪɛla***.

nããpɪɛlgantal [nã̀ã̀pɪ́ɛ́lgàntàl̀] *n.* top of the foot.

nããpɪɛlpatʃɪgɪɪ [nã̀ã̀pɪ́ɛ́lpàtʃɪ́gɪ́ɪ́] *n.* sole of the foot.

nããpol [nã̀ã̀pól] *n.* Achilles tendon. *pl.* ***nããpolo***.

naasaara [nààsáárá] (*var.* **nansaaraa**, **naasaarpʊmma**) *n.* Caucasian person, may also apply to non-Africans generally. (ultm. Arabic, via Hausa <*nasaara* 'Nazarenes (Christians)'). *pl.* ***naasarasa***.

naasaarbaal [nààsààrbáàl] *n.* white, Caucasian man. *pl.* ***naasaarbaala***.

naasaardaa [nààsààrdáá] *n.* Neem tree *syn:* **naasaarsɪŋtʃaʊ**; **naasaargbɛsa** (*Azadirachta indica*). *pl.* ***naasaardaasa***.

naasaargbɛsa [nààsààrgbɛ́sà] *n.* type of tree *syn:* **naasaardaa** .

naasaarhããŋ [nààsààrhã́ã̀ŋ] *n.* white, Caucasian woman. *pl.* ***naasaarhããna***.

naasaarlulii [nààsààrlúlíí] *n.* non-local medicine, such as pills and other packaged medicine.

naasaarpʊmma (*var. of* **naasaara**)

naasaarsɪŋtʃaʊ [nààsààrsɪ́ŋtʃáʊ́] *n.* Neem tree *syn:* **naasaargbɛsa**; **naasaardaa** .

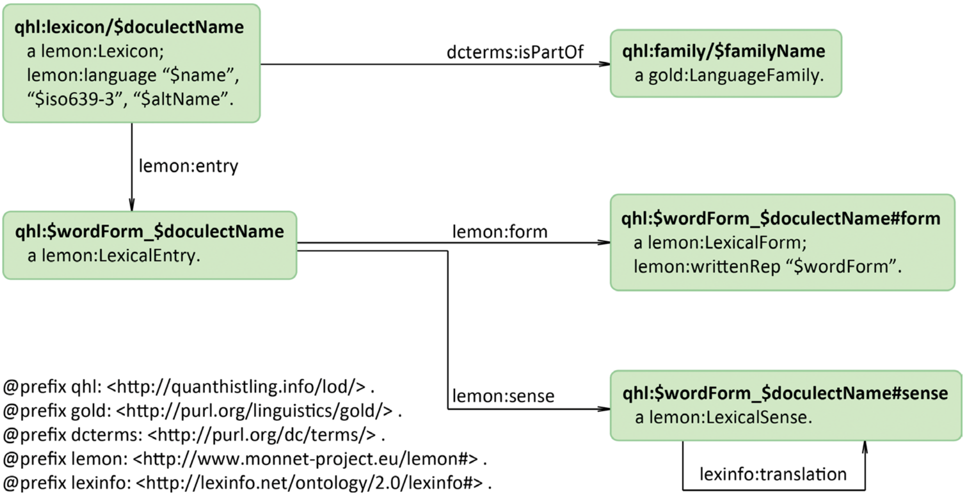
113

Figure 4.4

Wordlist and dictionary exemplars (above and opposite).

```
@date: 2012-11-23
@url: http://www.quanthistling.info/data/source/aguiar1994/dictionary-329-369.html
@source_title: Analise descritiva e teorica do Katukino-Pano
@source_author: de Aguiar, Maria Sueli
@source_year: 1994
@doculect: Katukina, n/a, Katukina, Panoan
@doculect: Portugues, por, Portugues, Panoan
QLCID HEAD HEADDOCULECT TRANSLATION TRANSLATIONDOCULECT
aguiar1994/329/1 ai Katukina presente Portugues
aguiar1994/329/2 aima Katukina solteiro Portugues
aguiar1994/329/3 ain Katukina esposa Portugues
aguiar1994/329/4 ainnan Katukina cipo para cesta Portugues
aguiar1994/329/5 ainnan Katukina casado Portugues
aguiar1994/329/6 aka Katukina soco Portugues
aguiar1994/329/7 akaai Katukina tomar Portugues
```

Figure 4.5
QuantHistLing data extraction format.



```
@prefix qhl: <http://quanthistling.info/lod/> .
@prefix gold: <http://purl.org/linguistics/gold/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix lexinfo: <http://lexinfo.net/ontology/2.0/lexinfo#> .
```

Figure 4.6
An implementation of QuantHistLing data modeled in *lemon*.

interoperable resource that we call a *translation graph*—an RDF model that allows users to query across the underlying lexicons and dictionaries to extract semantically aligned wordlists via their glosses and translations.[32] Identifying semantically related sets of words from different languages is one step in investigating the historical evolution of languages and their possible relatedness.[33]

Conversion of wordlist and dictionary data from QuantHistLing into *lemon* has the advantage that *lemon* is tightly integrated with Semantic Web technologies. In particular, lexical data in *lemon* are easily made interoperable with the Linguistic Linked Open Data

(LLOD) cloud. Thus, the resulting lexical resource is available on the web in a standard format and accessible, the data can be made query-able via a SPARQL endpoint,[34] and the use of the *lemon* ontology with Linked Data assists QuantHistLing in its goals to merge disparate dictionary and wordlist data via semantic sense and meaning mappings into an ontology for graph-to-CSV extraction of multilingual and disparate resources.[35]

This is indicative of researchers' efforts at transforming multilingual lexical datasets into Semantic Web data. That is, there exists some input data format (often CSV) from which lexical semantic data needs to be mapped to similar nodes in a given translation graph. Furthermore, metadata about languages or resources in the dataset must be annotated with URIs so that those resources can be linked to other datasets. This linking lies at the heart of the Linked Data initiative, and in particular of the LLOD, which aims to make available an increasing number of resources on under-resourced languages to research communities via the web.

PHOIBLE in CLLD

The PHOIBLE database is a broad collection of spoken languages' phonological systems.[36] It encodes a theory of linguistic description that includes systems of phonemes, allophones, and their phonological conditioning environments. The formalism is known as distinctive feature theory, is semi-binary, and has been used to model broad-base applications for automatic spoken-language (even dialect) recognition. Distinctive feature theory in phonology was developed in the early-to-mid-20th century as an abstraction of the physical acoustic signals (in speech) into a graphemic-based encoding (that is, letter-based transcription) of sounds and their contrasts. This theory allows linguists to describe and predict (un)natural classes of sound changes.

PHOIBLE was initially published as Linked Data in a simple RDF model, which includes concepts (languages, sounds, and features) and the relationsbetween languages and their sounds, and sounds and their features (Moran 2012a, 2012b). This prototype was created by scripting input in CSV data and outputting an RDF graph, given a model, into an XML serialization. More recently, the PHOIBLE data has been incorporated into the Cross-Linguistic Linked Data (CLLD) framework (Forkel 2014). For under-resourced languages, the CLLD framework provides several straightforward mechanisms for taking structured data (say, CSV and BibTeX for bibliographic references), especially from diverse linguistics datasets like typological databases,[37] and generating end-user-friendly interfaces with features like explorable maps, sortable features, and searchable content.[38]

Beyond just a nice web interface, CLLD applications provide their data as Linked Data described with VoID descriptions, and those data are accessible through tools such as rdflib[39] and Python.[40] The core CLLD data model is illustrated in figure 4.7, which contains concepts (Dataset, Language, Parameter, ValueSet, Value, Unit, UnitParameter, UnitValue, Source, Sentence, Contribution) and the relations between entities—providing a triples model (Forkel 2014).[41]

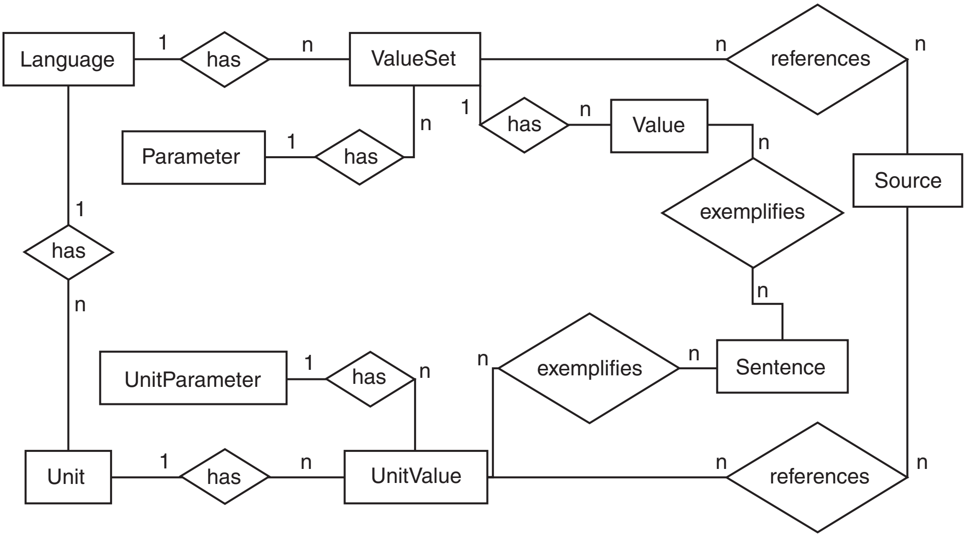


Figure 4.7
Entity-relationship diagram of the CLLD core data model.

The impact of CLLD applications is spelled out in Forkel (2014). In sum, queries like "give me all information on language X" are possible, and they will return all information from all CLLD applications for a given language. The query functionality also allows for testing conjectures made in particular sources, such as the WALS chapter "Hand and Arm" (Brown 2013), on the evolution of languages and other aspects of linguistic diversity. More complex queries that federate the CLLD resources are also possible via the CLLD Portal.[42] Extracted data can then be used either to seed or to expand the development of other datasets with language metadata, linguistic features, and lexical and orthographic encoded data—in particular, data on under-resourced languages that may be used in social media outlets such as social networks, blogs, or tweets.

Combining Case Studies

We have already presented two brief case studies of the transformation of linguistic data into Linked Data. Now we may ask, what can we do with these resulting Linked Data resources? One idea is that we might want to reconsider the notion of resource porting through character-based machine translation. For example, using the PHOIBLE vocabulary, we can describe languages on the level of their phonemic structure and, subsequently, we can also describe the systematic sound correspondences between different languages. We have an appropriate target dataset in QuantHistLing.

At the moment, character-based machine translation manages to identify corresponding characters or character groups, yet treats them as opaque signs. In fact, however, sound correspondences tend to reflect systematic laws, meaning that not one specific phoneme developed into another, but that *all phonemes* with a specific feature turned into phonemes whose feature value was replaced by another value. Unlike state-of-the-art character-based models, a phoneme-level model would be able to capture this information if a mapping from character to phoneme (or phonetic feature set) can be established.[43] This is, however, a direction for future research, and it requires a close integration of linguistic and NLP expertise. Under the umbrella of the interdisciplinary Open Linguistics Working Group (OWLG), however, such a collaboration may be possible, because it represents one of the very few forums where both communities actually meet.

The Linguistic Linked Open Data Cloud

Recent years have seen not only a number of approaches to provide linguistic data as Linked Data, but also the emergence of larger initiatives that aim at interconnecting these resources. Among these, the Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation (OKFN) has spearheaded the creation of new data and the republishing of existing linguistic resources as part of the emerging Linguistic Linked Open Data (LLOD) cloud. These initiatives provide technological infrastructure and community support for researchers wishing to produce and share under-resourced language data.

The LLOD Cloud

Aside from benefits arising from the actual *linking* of linguistic resources, various linguistic resources from very different fields have been provided in RDF and related standards over the last decade. In particular, this is the case for lexical resources like WordNet (Gangemi, Navigli, and Velardi 2003), which represents a cornerstone of the Semantic Web and is firmly integrated in the Linked Open Data (LOD) cloud. In a broader sense, LOD general knowledge bases from the LOD such as the DBpedia have also been rendered as lexical resources, owing to their immanent relevance for Natural Language Processing tasks such as Named Entity Recognition (NER) or Anaphora Resolution (AR). Other types of linguistically relevant resources with less importance to AI and knowledge representation, however, are not a traditional part of the LOD cloud, although they do motivate the creation of a sub-cloud dedicated to linguistic resources.

Figure 4.8 illustrates the Linguistic Linked Open Data (LLOD) cloud diagram. The LLOD cloud is a collection of linguistic resources that are published (typically) under open licenses as Linked Data. The data are decentralized, developed, and maintained with metadata online.[44] The cloud diagram is developed as a community effort in the context of OWLG and is built automatically from metadata about Linked Data sources stored online. Users who wish to have their datasets included need to make sure that at least one URL provided for data or endpoints is up and running. Metadata tags for discoverability include

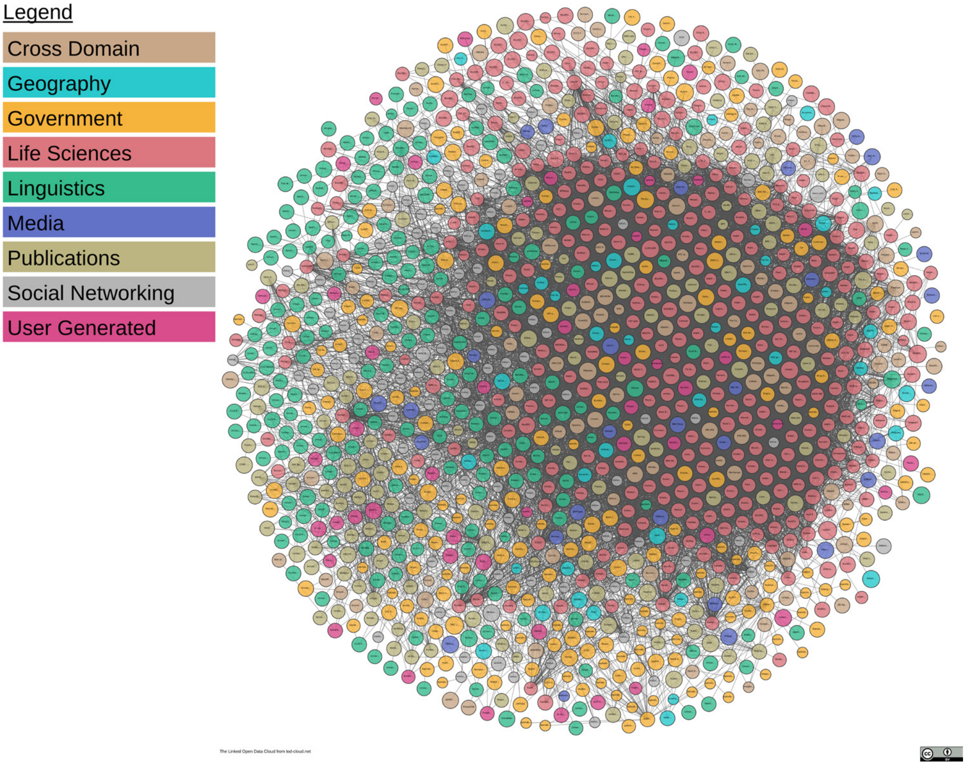


Figure 4.8
Linguistic Linked Open Data (LLOD) cloud.

“llod” and “linguistics.” Other tags are used to more precisely define specific resources (e.g., corpus, lexicon, wordnet, thesaurus).

The Open Linguistics Working Group

The LLOD cloud is a result of a coordinated effort by the Open Linguistics Working Group (OWLG; see Chiarcos and Pareja-Lora, this volume).

Since its formation in 2010, the OWLG has grown steadily. One of our primary goals is to attain openness in linguistics through:

1. Promoting the idea of open linguistic resources
2. Developing the means for the representation of Open Data
3. Encouraging the exchange of ideas across different disciplines

Publishing linguistic data under open licenses is an important issue in academic research, as well as in the development of applications. We see increasing support for this in the linguistics community (Pederson 2008), and there are a growing number of resources published under open licenses (Meyers et al. 2007). Publishing resources under open licenses offers many advantages: For instance, freely available data can be more easily reused, double investments can be avoided, and results can be replicated. Also, other researchers can build on the data and subsequently can refer to the publications associated with them. Nevertheless, a number of ethical, legal, and sociological problems are associated with Open Data,[45] and the technologies that establish interoperability (and thus reusability) of linguistic resources are still under development.

The OWLG represents an open forum for interested individuals to address these and related issues. At the time of writing, the group consists of about 100 people from 20 different countries. Our group is relatively small, but continuously growing and sufficiently heterogeneous. It includes people from library science, typology, historical linguistics, cognitive science, computational linguistics, and information technology; the ground for fruitful interdisciplinary discussions has been laid out. One concrete result emerging out of collaborations between a large number of OWLG members is the LLOD cloud, as already sketched above. Independent research activities of many community members involve the application of RDF/OWL to represent linguistic corpora, lexical-semantic resources, terminology repositories, and metadata collections about linguistic data collections and publications. To many such members, the Linked Open Data paradigm represents a particularly appealing set of technologies. Within the OWLG, these activities have converged toward building the cloud.

Under-Resourced Languages in the LLOD Cloud

Two principal driving forces of the growth of the LLOD cloud diagram and the OWLG have been, first, the synergies between independent research projects whose experts were interested in providing their data as RDF or Open Data, and, second, multinational projects, often funded by the EU, that focus on technological solutions for multilinguality issues in the European digital single market (affecting matters of localization, computational lexicography, and machine translation). A third factor that contributed to this development has been more recent projects and applications in the humanities and academic branches of linguistics. With the research described in this paper, we demonstrate the applicability of LLOD technologies to one of these “small” areas of research and their ability to harness their highly specific resources in studying under-resourced languages. We consider the adaptation of this technology in an area where both experts and students are often lacking programming skills to be a particularly strong case for the potential of Linked Data in linguistics.

However, the QuantHistLing projects and CLLD are only two exemplary case studies from this particular area. Related efforts that employ RDF and/or Linguistic Linked Open Data for the study and comparison of less-resourced languages include, for example, the “Typology Tool” TYTO (Schalley 2012) that utilizes Semantic Web technologies to process, integrate, and query cross-linguistic data. The Typological Database System[46] (Dimitriadis et al. 2009) uses OWL ontologies for harmonizing and providing access to distributed databases that are created in the course of typological research and language documentation. For a similar application in language resource harmonization, the GOLD ontology was created as part of the Electronic Metastructure for Endangered Languages Data (E-MELD, see Langendoen, this volume).

Poornima and Good (2010) have already described the application of RDF and Linked Data technologies for creating machine-readable wordlists for under-resourced languages. Building on these and other pieces of earlier research, the project called Linked Open Dictionaries (LiODi) is currently developing techniques to facilitate cross-linguistic search across dictionaries to assist in language contact studies among endangered and historical languages in the Caucasus area and among Turkic languages (Abromeit et al. 2016), as well as to assist in the LLOD conversion of formats typically used in linguistic typology and for language documentation (Chiarcos et al. 2017). While these technologies and the resources created on this basis are still under development, the PanLex project (Kamholz, Pool, and Colowick 2014) has already published a near-universal RDF-based translation graph that covers numerous under-resourced languages.

Getting Additional Guidance

As is the case when experts adopt any state-of-the-art technologies, advances and developments are happening faster than traditional print media can possibly keep up with. In this paper, we provided sound reasoning and examples of why we believe Linked Data is an important platform for working with and disseminating under-resourced language data. Nevertheless, the tools and technologies currently up to speed will have inevitably gained much ground before this volume makes it to press. Therefore, we have put together a repository where we store our recent educational materials and do-it-yourself tutorials for users who wish to implement and publish models of Linguistic Linked Open Data with their own resources.[47]

Summary

This chapter provides a general introduction to Linked Data and its application in the language sciences, with a specific emphasis on its uses for studying under-resourced languages. We identified characteristics of data for such languages, focusing on lexical resources (wordlists and dictionaries) and on annotated corpora (glosses and corpora). We further discussed aspects of resource integration, before focusing on Linked Data and under-resourced language data in particular. We then homed in on Linked Data for linguistics and NLP, and we gave two brief case studies of linguistic data sources that have been transformed into Linked Data. Finally, we described in detail the status and the bandwidth of applications of Linked Open Data technologies to under-resourced languages in the general context of the Open Linguistics Working Group and the developing Linguistic Linked Open Data (LLOD) ecosystem.

Notes

1. http://tei-c.org.

2. https://www.iso.org/developing-standards.html.

3. http://linguistics.okfn.org/.

4. https://www.w3.org/community/ontolex/.

5. http://www.lider-project.eu/.

6. Furthermore, increased access to language descriptions leads to increased documented typological diversity (at least in phonology, cf. Moran 2012a).

7. http://www.endangeredlanguages.com.

8. Important language catalogs include the Ethnologue (Lewis, Simons, and Fennig 2014) and the Open Languages Archive Network (OLAC).

9. http://glottolog.org.

10. Any concrete definition of the "under-resourced-ness" of languages' data should probably include a checklist of data types, as in "language X has a grammar, a dictionary, a corpus, a treebank." This definition would be problematic because what we know about worldwide language documentation is dynamic. Not only is documentation increasing, it is also decreasing, as for instance when the last records of language X are encoded in no longer accessible (electronic) formats.

11. Even more frightening for linguists studying linguistic diversity is that around one-third of the currently spoken languages are believed to be language isolates, or languages that are the last remaining leaf node in their language family tree. When lost, these languages take with them any typological structures that may not be accounted for anywhere else in the world. This phenomenon has often been compared to the loss of a biological species, which thereby limits biologists' view and study of the evolutionary processes that lead to worldwide diversity.

12. http://www.ruscorpora.ru/en/.

13. Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene.

14. Icelandic, Irish, Latvian, Lithuanian, Maltese.

15. QuantHistLing is a project that has extracted wordlist data from many resources and uses both Linked Data and the ontological model called Lexicon Model for Ontologies (*lemon*; McCrae et al. 2010, 2011) (http://lemon-model.net/) to combine data sources.

16. http://corpora.informatik.uni-leipzig.de/.

17. Numerous examples: http://odin.linguistlist.org.

18. http://www-01.sil.org/computing/toolbox/.

19. http://dingo.sbs.arizona.edu/~sandiway/treebanksearch/.

20. http://www.mlta.uzh.ch/en/Projekte/Baumbanken.html.

21. The term "resource" is ambiguous: *Linguistic* resources are structured collections of data that can be represented, for example, in RDF. In RDF, however, "resource" is the conventional name of a node

in the graph, because, historically, these nodes were meant to represent objects described by metadata. In ambiguous cases, we use the terms "node" or "concept" whenever *RDF* resources are meant.

22. One example is the General Ontology of Linguistic Description (GOLD) by Farrar and Langendoen (2003).

23. For example, structured output from frameworks like Head-driven Phrase Structure Grammar (HPSG) or Lexical Functional Grammar (LFG).

24. http://clld.org.

25. http://www.lexvo.org/.

26. http://glottolog.org.

27. http://phoible.org.

28. http://panlex.org/.

29. http://www.acoli.informatik.uni-frankfurt.de/liodi/.

30. QuantHistLing was funded from 2010 to 2014 by the European Research Council (Michael Cysouw, University of Marburg, primary investigator). Its aims were to uncover and clarify phylogenetic relationships between native South American languages, particularly the Tukonoan, Witotoan, and Jivoroan language families, using quantitative methods. The two main objectives were the digitalization of the lexical resources on native South American languages and the development of innovative computer-assisted methods to quantitatively analyze this information. The project focused on formalizing (i.e., computationally coding) aspects both of data transformation and of the comparative method, by collaborating with research scientists in other fields.

31. Data are online at http://cysouw.de/home/quanthistling.html.

32. For a broad application of a translation graph aimed at worldwide coverage, see PanLex (Kamholz, Pool, and Colowick 2014): http://panlex.org.

33. Another necessary step is the identification of cognates via shared sound correspondences—a signal of genealogical relatedness. This process is comparable to DNA string comparison algorithms from bioinformatics, which have been reapplied and recoded for linguistic purposes (cf. List and Moran 2013).

34. There is an endpoint at http://www.linked-data.org:8890/sparql.

35. QuantHistLing data available in RDF and *lemon*: http://www.linked-data.org/datasets/qhl.ttl.zip.

36. http://phoible.org.

37. http://clld.org/datasets.html.

38. CLLD applications can conveniently use the Github "pull" functionality; in other words, CLLD project-specific applications can retrieve data directly from online hosted data and code repositories.

39. https://github.com/RDFLib/rdflib.

40. http://nbviewer.ipython.org/gist/xflr6/9050337/glottolog.ipynb.

41. There are several RDF serialization formats (e.g., Turtle, N-triples, XML). We do not go into detail with regard to them here.

42. Full SPARQL functionality is not supported. See: http://portal.clld.org/.

43. See Moran and Cysouw (2018) for a systematic exposition.

44. Originally, LLOD metadata was maintained under http://datahub.io. At the time of writing, LLOD metadata is being maintained under http://linghub.org. Because the LLOD cloud diagram is

now generated as a view of the LOD cloud diagram, novel datasets can be added via https://lod-cloud.net/add-dataset.

45. For example, complex copyright situations may arise if one resource (say, a lexicon) were to be developed on the basis of a second resource (say, a newspaper archive) and researchers felt uncertain whether the examples from the original newspaper contained in the lexicon violate the original copyright. Ethical problems may arise if a database of quotations from a newspaper were linked to a database of speakers and that database were further connected with, say, obituaries from the same newspaper. Even if this were done only in order to study generation-specific language variation, one may wonder whether such an accumulation of information violates the privacy of the people involved.

46. https://languagelink.let.uu.nl/tds/.

47. http://acoli.informatik.uni-frankfurt.de/resources/llod/index.html.

References

Abromeit, F., C. Chiarcos, C. Fäth, and M. Ionov. 2016. "Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF." In *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*, 11–19. Portoroz, Slovenia, ELRA.

Benson, M. 1990. "Collocations and General-Purpose Dictionaries." *International Journal of Lexicography* 3 (1): 23–34.

Bickel, B., and J. Nichols. 2015. Autotyp. http://www.autotyp.uzh.ch/.

Brown, C. H. 2013. "Hand and Arm." In *The World Atlas of Language Structures Online,* edited by M. S. Dryer and M. Haspelmath. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Carlson, L., M. E. Okurowski, D. Marcu, L. D. Consortium et al. 2002. RST Discourse Treebank. Linguistic Data Consortium, University of Pennsylvania.

Chiarcos, C. 2008. "An Ontology of Linguistic Annotations." *LDV Forum* 23 (1): 1–6.

Chiarcos, C., M. Ionov, M. Rind-Pawlowski, C. Fäth, J. W. Schreur, and I. Nevskaya. 2017. "LLOD-ifying Linguistic Glosses." In *International Conference on Language, Data and Knowledge*, 89–103. Galway, Ireland. Springer: Cham.

Chiarcos, C., J. McCrae, P. Cimiano, and C. Fellbaum. 2013. "Towards Open Data for Linguistics: Linguistic Linked Data." In *New Trends of Research in Ontologies and Lexical Resources*, edited by A. Oltramari, Lu-Qin, P. Vossen, and E. Hovy. Heidelberg: Springer.

Chiarcos, C., S. Nordhoff, and S. Hellmann. 2012. *Linked Data in Linguistics*. Berlin, Heidelberg: Springer.

Clark, A. 2003. "Combining Distributional and Morphological Information for Part of Speech Induction." In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, 59–66. Association for Computational Linguistics.

Cysouw, M. 2005. "Quantitative Methods in Typology." In *Quantitative Linguistics: An International Handbook*, edited by G. Altmann, R. Köhler, and R. G. Piotrowski, 554–578. Berlin: Walter de Gruyter.

de Melo, G. 2015. "Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud." *Semantic Web Journal* 6 (4): 393–400.

Dimitriadis, A., M. Windhouwer, A. Saulwick, R. Goedemans, and T. Bíró. 2009. *How to Integrate Databases without Starting a Typology War: The Typological Database System. The Use of Databases in Cross-Linguistic Studies*, 155–207. Berlin: Mouton de Gruyter.

Donohue, M., R. Hetherington, J. McElvenny, and V. Dawson. 2013. World phonotactics database. Department of Linguistics, Australian National University. http://phonotactics.anu.edu.au.

Dryer, M. S., and M. Haspelmath. 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Evans, N., and S. C. Levinson. 2009. "The Myth of Language Universals: Language Diversity and Its Importance for Cognitive Science." *Behavioral and Brain Sciences* 32:429–448.

Farrar, S., and T. Langendoen. 2003. "A Linguistic Ontology for the Semantic Web." *GLOT* 7 (3): 97–100.

Forkel, R. 2014. "The Cross-Linguistic Linked Data Project." In *Proceedings of the Third Workshop on Linked Data in Linguistics (LDL 2014)*, 60–66. Reykjavik, Iceland, ELRA.

Gangemi, A., R. Navigli, and P. Velardi. 2003. "The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet." In *Proceedings of On the Move to Meaningful Internet Systems (OTM2003)*, edited by R. Meersman and Z. Tari, 820–838. Catania, Italy.

Good, J. 2013. "Fine-Grained Typological Investigation of Grammatical Constructions Using Linked Data." In *Proceedings of the Tenth Biennial Conference of the Association of Linguistic Typology (ALT X),* Leipzig.

Hammarström, H., R. Forkel, M. Haspelmath, and S. Bank. 2015. Glottolog 2.6. Jena: Max Planck Institute for the Science of Human History. http://glottolog.org.

Ide, N., and J. Pustejovsky. 2010. "What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability." In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong.

Ide, N., and K. Suderman. 2007. "GrAF: A Graph-Based Format for Linguistic Annotations." In *Proceedings of the 1st Linguistic Annotation Workshop (LAW 2007)*, Prague, Czech Republic. Association of Computational Linguistics.

ISO 639-3:2007. Codes for the representation of names of languages—Part 3: Alpha-3 code for comprehensive coverage of languages. Geneva: International Organization for Standardization.

Jones, D., and R. Havrilla. 1998. "Twisted Pair Grammar: Support for Rapid Development of Machine Translation for Low Density Languages." In *Machine Translation and the Information Soup*, edited by D. Farwell, E. Hovy, and L. Gerber, 318–332. Berlin: Springer.

Kamholz, D., J. Pool, and S. M. Colowick. 2014. "PanLex: Building a Resource for Panlingual Lexical Translation." In *Proceedings of the Ninth Language Resources and Evaluation Conference (LREC 2014)*, 3145–3150. Reykjavik, Iceland, ELRA.

Kingsbury, P., and M. Palmer. 2002. "From TreeBank to PropBank." In *Proceedings of the Third Language Resources and Evaluation Conference (LREC 2002)*, 1989–1993. Las Palmas de Gran Canaria, Canary Islands, Spain ELRA.

Kučera, H., and W. N. Francis. 1967. *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Lewis, M. P., G. F. Simons, and C. D. Fennig. 2014. *Ethnologue: Languages of the World, Seventeenth edition*. Dallas: SIL International.

List, J.-M., and S. Moran. 2013. "An Open Source Toolkit for Quantitative Historical linguistics." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, (ACL 2013), 13–18. Sofia, Bulgaria, Association of Computational Linguistics.

Matisoff, J. A. 2015. Sino-tibetan etymological dictionary and thesaurus (stedt). http://stedt.berkeley.edu/.

Maxwell, M., and B. Hughes. 2006. "Frontiers in Linguistic Annotation for Lower-Density Languages." In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, 29–37. Sydney, Australia,.Association of Computational Linguistics.

McCrae, J., G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. G. Pérez, J. Gracia, et al. 2010. *The Lemon Cookbook*. Technical report, CITEC, Universität Bielefeld, Germany.

McCrae, J., D. Spohr, and P. Cimiano. 2011. "Linking Lexical Resources and Ontologies on the Semantic Web with Lemon." In *The Semantic Web: Research and Applications*, *Proceedings of the 2nd European Semantic Web Conference (LNCS 3532)*, 245–259. Springer.

McNew, G., C. Derungs, and S. Moran. 2018. "Towards Faithfully Visualizing Global Linguistic Diversity." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 805–809. May 7–12, Miyazaki, Japan. http://www.lrec-conf.org/proceedings/lrec2018/pdf/813.pdf.

Meyers, A., N. Ide, L. Denoyer, and Y. Shinyama. 2007. "The Shared Corpora Working Group Report." In *Proceedings of the First Linguistic Annotation Workshop (LAW-I), held in conjunction with ACL-2007,* 184–190. Prague, Czech Republic. Association of Computational Linguistics.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. "Annotating Noun Argument Structure for NomBank." In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2004)*, 803–806. Lisbon, Portugal, ELRA.

Moran, S. 2009. "An Ontology for Accessing Transcription Systems (OATS)." In *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009)*, Athens, Greece. Association for Computational Linguistics.

Moran, S. 2012a. "Phonetics Information Base and Lexicon." PhD diss., University of Washington.

Moran, S. 2012b. "Using Linked Data to Create a Typological Knowledge Base." In *Linked Data in Linguistics*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann,129–138. Berlin: Springer.

Moran, S., and M. Brümmer. 2013. "Lemon-Aid: Using Lemon to Aid Quantitative Historical Linguistic Analysis." In *Proceedings of the Second Workshop on Linked Data in Linguistics: Representing and Linking Lexicons, Terminologies and Other Language Data*, 28–33. Pisa, Italy, Association of Computational Linguistics.

Moran, S., and M. Cysouw. 2018. "The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles." *Translation and Multilingual Natural Language Processing series in Language Science Press.* DOI: https://doi.org/10.5281/zenodo.1296780; http://langsci-press.org/catalog/book/176.

Moran, S., D. McCloy, and R. Wright. 2014. *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Nordhoff, S., H. Hammarström, R. Forkel, and M. H., eds. 2013. Glottolog 2.2. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://glottolog.org.

Pederson, T. 2008. "Empiricism Is Not a Matter of Faith." *Computational Linguistics* 34 (3): 465–470.

Poornima, S., and Good, J. 2010. "Modeling and Encoding Traditional Wordlists for Machine Applications." In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, 1–9. Uppsala, Sweden, Association for Computational Linguistics.

Pradhan, S. S., L. Ramshaw, R. Weischedel, J. MacBride, and L. Micciulla. 2007. "Unrestricted Coreference: Identifying Entities and Events in OntoNotes." 1st *IEEE International Conference on Semantic Computing (ICSC)*, 446–453. Irvine, CA, IEEE.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber. 2008. "The Penn Discourse TreeBank 2.0." In *Proceedings of the Sixth Language Resource and Evaluation Conference (LREC 2008)*, 2961–2968. Marrakesh, Morocco.

Prud'hommeaux, E., and A. Seaborne. 2008. SPARQL Query Language for RDF. W3C Recommendation January 15, 2008.

Pustejovsky, J., P. Hanks, R. Sauri, A. See R. Gaizauskas, A. Setzer, et al. 2003. "The TimeBank Corpus." In *Proceedings of Corpus Linguistics 2003. UCREL technical paper number 16*, 647–656. UCREL, Lancaster University, UK.

Rehm, G., and H. Uszkoreit. 2013. *META-NET Strategic Research Agenda for Multilingual Europe 2020*. Berlin: Springer.

Schalley, A. C. 2012. "TYTO—A Collaborative Research Tool for Linked Linguistic Data." In *Linked Data in Linguistics*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 139–149. Berlin: Springer.

Snyder, B., R. Barzilay, and K. Knight. 2010. "A Statistical Model for Lost Language Decipherment." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1048–1057. Uppsala, Sweden, Association for Computational Linguistics.

Taylor, A., M. Marcus, and B. Santorini. 2003. "The Penn Treebank: An Overview." In *Treebanks (Text, Speech and Language Technology)*, edited by A. Abeillé, vol. 20, 5–22. Dordrecht: Springer.

Tiedemann, J. 2012. "Character-Based Pivot Translation for Under-Resourced Languages and Domains." In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 141–151 (EACL 2012). Avignon, France: Association for Computational Linguistics.

Windhouwer, M., and S. Wright. 2012. "Linking to Linguistic Data Categories in ISOcat." In *Linked Data in Linguistics*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 99–107. Berlin: Springer.

Wright, S. 2004. "A Global Data Category Registry for Interoperable Language Resources." In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2004)*, 123–126. Lisboa, Portugal.

Yarowsky, D., G. Ngai, and R. Wicentowski. 2001. "Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora." In *Proceedings of the First International Conference on Human Language Technology Research*, 1–8. San Diego, CA, Association of Computational Linguistics.

Zock, M., and S. Bilac. 2004. "Word Lookup on the Basis of Associations: From an Idea to a Roadmap." In *COLING 2004: Enhancing and Using Electronic Dictionaries*, edited by M. Zock, 29–35. Geneva, Switzerland: Association of Computational Linguistics.

5

A Data Category Repository for Language Resources

Kara Warburton and Sue Ellen Wright

Language Resources

DatCatInfo is an online resource of information about data categories (DCs) that are used in natural language processing applications and in the research and development of language resources. The collection was originally called "the Data Category Registry" and commonly referenced as the "DCR." Maintained under the web address https://www.isocat.org, this collection was developed by the International Organization for Standardization Technical Committee 37 for *Language and Terminology* (ISO/TC 37) and was configured as a standardized ISO Registry, with the Max Planck Institute for Psycholinguistics (Nijmegen, the Netherlands; hereinafter designated as MPI) acting as an official Registration Authority.

Over time it became clear that although a wide range of linguists were interested in documenting data categories, few supported standardizing them by following a three-stage balloting procedure prescribed by ISO. Consequently, the "Registry" has been rechristened a "Data Category Repository," and it has been undergoing a major refit since 2014. In this article and other publications about DatCatInfo and its history, the term "Registry" is thus reserved for the ISO sponsored resource (up to 2014), and "Repository" is used for DatCatInfo (after 2014). The well-established acronym "DCR," once used for "Data Category Registry" in contexts describing "DatCatInfo," now refers to "Data Category Repository."

The main purpose of DatCatInfo is to support the development of language resources, and yet DatCatInfo is itself a language resource. This chapter will therefore start with a brief discussion about language resources. According to the European Language Resource Association (ELRA), the term *language resource* refers to "a set of speech or language data and descriptions in machine readable form," such as electronic corpora, terminology databases (termbases), and computational lexicons (ELRA 2017). These resources are used to support a wide range of applications that are strategic for the digital economy, such as speech recognition and synthesis, knowledge mining, search engine optimization, content analysis and management, focused marketing, machine translation, and the language services industry at large (translation, interpreting, and localization). Language resources are pervasive: They are a core component of computer operating systems, they are essential for productivity applications such as office suites (word processing, spreadsheets, and the like), they are used in many kinds of automated industrial and commercial equipment, and they are even found in cell phones. When we make an online purchase, use an automated telephone service, or send a text message, we are using language resources. So-called artificial intelligence (AI) applications are built in part on extensive language resources.

Language resources are created and managed by translators, terminologists, lexicographers, linguists, researchers, software engineers, and numerous other professionals, many using specialized computational software. They are developed in academic research settings, in commercial environments, and in public institutions as well.

First developed on paper, and refined over time in digital environments, language resources have evolved in parallel with language industry standards. For instance, text corpora have become more powerful and useful as language resources with the evolution of the annotation framework standards produced in ISO/TC 37, and termbases are based on models described in theoretical standards such as ISO 704, as well as in practical standards such as ISO 30042, the TermBase eXchange standard (TBX). Without standards, developing language resources would be much more expensive than necessary, many steps and tasks would need to be duplicated that could otherwise be carried out only once, and it would be impossible to leverage information across language resources and across different applications. In short, lacking standards, our society would not have the range of language resources, and the applications that are enabled by them, that we have today.

Data Categories

One of the key areas of standardization that applies to language resources relates to their internal data structures—what kinds of data they contain and how these data are arranged. For example, before a spell checker can determine whether a word is correctly spelled or not, it needs to "know" if the word is a noun, verb, adjective, or other part of speech. "I *advise* [verb] my friend, but I give my friend *advice* [noun]." The difference—*s* or *c*—depends on the part of speech. Part-of-speech information is also crucial for semantic-based resources, such as electronic dictionaries and ontologies—since, despite the similarity in meaning, the definition of the verb will not be identical to that of the noun. The way words are categorized by their part of speech can be standardized, making spell checkers, as well as the many other language resources that use part of speech information, more interoperable.

The *part of speech* is an example of a linguistic *data category* (DC). There are hundreds, if not thousands, of different DCs that are found in language resources or are used to describe and manage concepts, names, data structures, and procedures common to language resources. There are also different types of DCs and a wide range of possible ways to document and describe them.

The Initial Data Category Selection: Terminology

In the 1990s, linguists and researchers, but in particular terminologists who were designing terminology management systems, began to share knowledge about DCs, with an aim to harmonize approaches and methodologies. In this context Wright and Budin conducted a study that documented DCs in all the then-available terminology management systems—which were legion at the time, but most of which are now defunct, as well as some national term banks, including Termium, Danterm, and even the old Sovterm (Wright and Budin 1994). Eventually, ISO TC 37, then charged primarily with developing standards for the field of terminology management, initiated harmonization efforts, which in turn led the ISO TC in 1999 to publish specifications for 215 DCs in the standard: *ISO 12620—Computer applications in terminology—Data categories*. In a sense, this standard comprised the first instantiation of DatCatInfo.

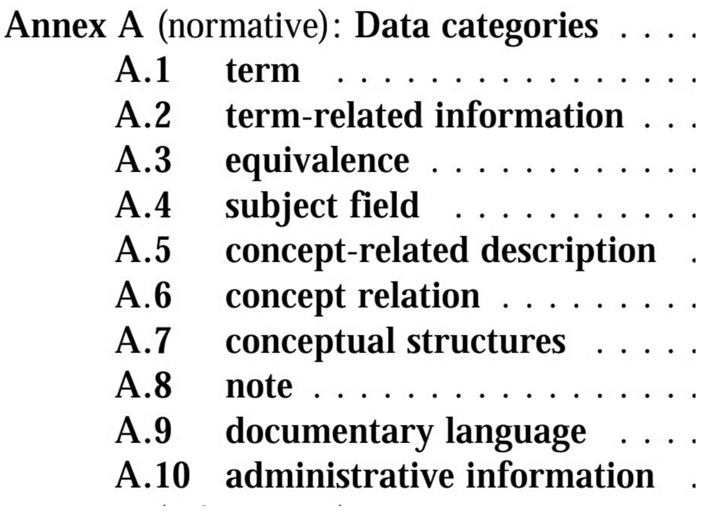
ISO 12620:1999 introduced the term *data category specification*, which is the sum of information that describes a DC. The structure and content of a data category specification as documented in that standard is shown in figure 5.1.

The DCs in 12620:1999 were grouped thematically, as shown in the table of contents (figure 5.2).

This thematic organization, in particular the division between *concept* and *term*, roughly mirrors the structural levels of a terminological entry as specified in ISO 12200, *Machine-Readable Terminology Interchange Format (MARTIF)*, a structural model that in 2003 was standardized in *ISO 16642–Terminological Markup Framework (TMF)*. The markup for termbases described in ISO 12200 was serialized in Standard Generalized Markup Language (ISO 8879:1986), which later gave rise to XML. ISO 12200 also aligns with the

Specification category	**Representation**
Notation number:	**boldface number**
Preferred data category name:	**boldface**
Admitted name:	ADMITTED NAME: **boldface** [repeatable]
Full form:	FULL FORM: **boldface**
Related name:	RELATED NAME: **boldface** [repeatable]
Nonadmitted name:	NONADMITTED NAME: **boldface** [repeatable]
Data category description:	DESCRIPTION:
Note:	NOTE: [repeatable]
Permissible instances:	PERMISSIBLE INSTANCES: *italics*
Example:	EXAMPLE: [repeatable]

Figure 5.1
Elements of a data category specification, from ISO 12620:1999.



Annex A (normative): **Data categories** . . . .
A.1 term . . . .
A.2 term-related information . . .
A.3 equivalence . . . .
A.4 subject field . . . .
A.5 concept-related description .
A.6 concept relation . . . .
A.7 conceptual structures . . . . .
A.8 note . . . .
A.9 documentary language . . . .
A.10 administrative information .

Figure 5.2
Groups of DCs from 12620:1999.

so-called meta-elements (<descrip>, <termNote>, and <admin>) in TBX, the XML markup language for terminology (replacing the SGML of ISO 12200), which in 2008 was published as *ISO 30042*. Thus, the elaboration of data categories has proceeded in parallel with the development of other resources (like the World Wide Web) and other standards that are familiar today.

Some DCs take free text as their content, such as /definition/ (Clause A.5.1 in ISO 12620),[1] while the content of others is confined to a closed set of permissible values, such as /grammatical gender/ (A.2.2.2), which can contain only the values *masculine, feminine, neuter*, or *other*. The type of content that a DC can take is referred to as its *content model*. ISO 12620:1999 did not clearly distinguish DCs according to their content models. Some of the permissible values were treated as DCs themselves with full data category specifications (for instance, the 19 values of /term type/ in Clauses A.2.1.1–A.2.1.19), while others were merely listed in the data category specification of their "parent" DC—for instance, the aforementioned values of /grammatical gender/ are listed as (a), (b), (c), and (d) in A.2.2.2. However, at that time, ISO 12620 was distributed in paper format only, so this descriptive approach, although occasionally inconsistent, did not cause serious application problems.

ISO 12620:1999 also introduced the concept of a *data category selection* (DCS). Since no termbase would contain all 215 DCs, it was understood that terminologists would select those DCs that were necessary for the purpose and users of their termbases. Different selections of DCs would be required for different types of termbases, as, for instance, a government-sponsored term bank documenting a nation's official languages versus a corporate termbase designed to support global marketing. Some such selections could become recognized as a best practice for certain applications and purposes. Each selection of DCs was referred to as a DCS.

ISO 12620:1999 marked a major milestone in the development of terminology resources and of terminology management as a practice. It enabled terminologists to begin harmonizing their termbases, thus rendering them more interoperable and repurposable, and, as previously mentioned, it also acted as a catalyst for the development of other standards. The concept of *harmonization* implies the use of uniform DC names and industry agreement on DC definitions or descriptions, two features that are essential in order to exchange data among different termbases.

Proposal for a Data Category Registry

When ISO 12620:1999 was due for systematic review five years later, ISO TC 37 decided that a major change was necessary. At around the same time, the sub-committee TC 37/SC 4, "Language Resource Management," was created, bringing stakeholders in fields of language resource management beyond terminology (such as lexicology, morphology, annotation schemes, and corpus management) into the TC 37 community, along with the new types of language resources that these stakeholders develop. The original number of DCs—215—needed to increase significantly to accommodate their needs. Furthermore, distributing information about DCs in a paper document that was updated only once every six years at best, and that sold for over $300 USD, was not acceptable to the majority of users, who ideally required DCs in the form of accessible data that could be combined, subsetted, and manipulated in a variety of applications. DC specifications needed to be treated as discrete units of information—mini-documents themselves that are more conducive to ad-hoc lookup and subject to frequent additions and updates—similar to items in an online catalog.

It was therefore proposed that an electronic version of the data category specifications be created in the form of an online database. The idea of moving data category specifications from paper to electronic format was reinforced by the widespread desire to create a collaborative, web-based environment where developers and researchers in linguistics and related disciplines could document the types of data that they work with, which ultimately would increase interoperability, reduce duplication and redundancy, and foster research and innovation. As would be discovered later, however, this type of environment, in which the linguistic community at large would be participating on a frequent basis, could not strictly adhere to the fixed standardization model dictated by ISO. The key distinction to be made here is *standardization* versus *harmonization*: agreement on names and content without formal balloting for each and every data category specification.

At the same time that ISO/TC 37 was designing its electronic resource for data category specifications, ISO Central Secretariat (CS) was planning a similar initiative for other standards, which it termed the Concept Database (ISO/CDB). The CDB would have allowed online lookup of a variety of data objects found in published ISO standards, including terms and definitions, graphical symbols, codes (language, country, currency, etc.), units of measurement, product properties, and items in data dictionaries. Although a pilot version was launched in 2009, the project was supplanted by the currently available ISO Online Browsing Platform (ISO 2018; Kemps-Snijders et al. 2009). In the shadow of CDB development, the future TC 37 DC database was to be called the Data Category Registry (DCR), since it was envisioned that data category specifications would, as noted above, become "standardized" and "registered" under ISO's Registration Authority (RA) model.

12620:2009

To create the DCR, it was necessary to first define the data model and governance procedures. Over a period of several years, ISO/TC 37 elaborated the necessary framework, published as ISO 12620:2009—*Specification of data categories and management of a Data Category Registry for language resources*. It is important to note that this version of 12620 contains no actual DC specifications. Rather, it outlines the data model and the methodology for creating and managing the future DCR.

12620:2009 was elaborated in close collaboration with representatives from ISO Central Secretariat, so as to ensure that the DCR would support the standardization of DCs in accordance with the CDB. As previously noted, the Max Planck Institute for Psycholinguistics (MPI) was appointed by the ISO Technical Management Board to be the Registration Authority. 12620:2009 covered the following topics, among others:

- The role of data categories (DCs) in language resource management
- Data category selections (DCS)
- Requirements for a DCR
- Registration authority
- The data model of a DCR and its data category specifications
- Management procedures
- The data category interchange format (DCIF)

The DCR was officially launched in 2008 under the brand name "ISOcat."

A Typology of Data Categories

Because DC specifications were now provided in electronic form only, their structure needed to be extremely rigorous, and the consistent declaration of DC values became essential. It was decided to consider the value of a DC, such as /feminine/ for /grammatical gender/, to be a data category in its own right and to be classified as a *simple* DC. All other DCs are deemed to be *complex* because, unlike simple DCs, they have a *conceptual domain*, that is, a "set of valid value meanings" (12620:2009, Clause 3.1.5). Valid value meanings are either (a) very open in nature, such as for the DC /definition/, which can contain free text, or (b) constrained by a rule, such as for /date/, which follows a certain format, or (c) strictly constrained to only a closed set of enumerated (permissible) values, that is, expressed by simple DCs, such as the values /noun/, /verb/, /adjective/ for the closed DC /part of speech/.

The following typology of DCs based on their content model was elaborated. In this typology, the conceptual domain is a key differentiating factor:

- Complex DC: DC that has a conceptual domain
    - Open DC: complex DC whose conceptual domain is not restricted to an enumerated set of values, such as a /definition/
    - Constrained DC: complex DC whose conceptual domain is non-enumerated, but is restricted to a constraint specified in a schema-specific language or languages, such as a /date/ or a range of dates
    - Closed DC: complex DC whose conceptual domain is restricted to a set of enumerated simple data categories, such as /part of speech/
- Simple DC: DC that does not have a conceptual domain, but is itself a member of a one, such as /noun/, or /verb/ for /part of speech/

Differentiating DCs based on the conceptual domain would guide the design of the data model of the future DCR.

An Elaborate Data Model

To accommodate both the new digital format and the standardization and registration workflows, the data category specification model in 12620:2009 needed to be more rigorous compared to its 1999 predecessor. Each DC specification comprised multiple nested sections, each of which included multiple fields for metadata:

- Administration Information Section
- Description Section
    - Data Element Name Section
    - Language Section
        - Name Section
        - Definition Section
        - Example Section
        - Explanation Section
- Conceptual Domain
- Linguistic Section

  - Conceptual Domain
  - Example Section
- Explanation Section

A detailed description of the purpose and content of these sections is included in ISO 12620:2009. However, the following observations should be noted:

- The Administration Information Section was designed to handle the information necessary for the ISO-specific standardization and registration workflows, such as submitting, reviewing, approving, and standardizing DCs.
- The Description Section contains information pertaining to the DC as a whole:
  - The Data Element Name section contains machine-readable names of the DC, for example, what the DC is called in various representation schemes (for instance, *definition*, *part of speech*, and so on).
  - The repeatable Language Section contains information about this data category in specific languages—for instance, a name, definition, example, and explanation in English, another in German, and so forth.
- The Conceptual Domain specifies a DC's permissible content, for instance, for /part of speech/, /noun/, /verb/, /adjective/, /adverb/.
- The Linguistic Section is used to "specify the behavior of a complex data category in a specific object language" (Clause 7.7). For example, for the /part of speech/ DC, this section can be used to specify which part of speech values are relevant in Cantonese; in the case of /grammatical gender/, Spanish would only need /masculine/ and /feminine/, whereas German requires the addition of /neuter/.

Initial Years of the DCR

The Max Planck Institute (MPI) acted not only as the DCR's Registration Authority (RA); it also took on an even larger role in technical development, maintenance, hosting, and even promotion, under the direction of the DCR Board. The board comprised members of ISO/TC 37/SC 3, chaired by Dr. Sue Ellen Wright, a co-author of this chapter. MPI provided these services from launch in 2008 until the end of 2014.

Researchers associated with the MPI began populating the DCR with data category specifications. The data category collection that had been assembled in TC 37 and during a number of previous research efforts (notably the SALT project, DXLT Specification, 2000) and that had supported the original ISO 12620, was imported into the new ISOcat environment more or less intact, with greater attention paid to the collection of additional data than to the refinement of existing resources. Under the auspices of the CLARIN project (CLARIN 2017; see Trippel and Zinn, this volume) researchers interested in defining concepts used in data mining and information retrieval across linked linguistic resources began to document DCs that reflected interests ranging from basic semantic and syntactic analysis to highly specialized collections, such as a new Polish national termbase. Broad *profiles* (a type of linguistic categorization used in the DCR) that apply to a wide range of different language resources, such as Morphosyntax and Metadata, evolved in the DCR in parallel with the historical domain-specific profiles of Terminology and Lexicography. More specific profiles, such as Sign Language, also took shape. The DCR gradually invited participants from a variety of linguistic communities to share their data, which resulted in the addition of, for instance, the GOLD ontology categories (see Langendoen, this volume) and other similar collections into the DCR.

Thus, during those first six years, the DCR experienced impressive growth. Starting with the 215 DCs from 12620:1999, it grew to include a formidable 6,185 DCs from a dozen linguistic disciplines. More than 150 experts from nearly 80 organizations contributed. Largely due to this collaborative approach, the ISO-inspired standardization workflow that had been incorporated into the design was never used. One effort to test that workflow was a dismal failure. With no dedicated staff to act as gatekeepers of the DCR, quality could not be assured, and duplication, redundancy, and other problems began to occur. Some of the thematic areas of the DCR were well documented, while others were neglected, resulting in imbalance. The DCR suffered from growing too fast without dedicated resources.

Additionally, in many cases the vigor applied to the project was not always met with commensurate rigor, resulting in considerable differences in quality among the entries. The original TC 37 set of DCs had been elaborated by trained terminologists, who tended to follow strict rules for writing definitions and careful procedures for elaborating data category specifications. Some sets of DCs were originally elaborated in other languages and then translated (not always well) into the base language, which was English, while other entries were translated into numerous languages, with varying degrees of success. The emerging collection demonstrated significant inconsistencies in quality and intention, primarily as the result of a quasi-cloud-sourcing environment that, at least in part, lacked firm management.

Withdrawal of the Max Planck Institute and Selection of Termweb

The original premise of the DCR was that DCs are fixed; for instance, a *noun* is a noun, no matter where it occurs in a language resource. Over time it became clear, however, that different communities of practice needed to use DCs in different ways. Although actual termbase development is often messy and frequently fails to adhere to ideal practice, terminologists designing the original DCR wanted to use DCs as clearly defined field names in their termbases in a way that would support reliable data interchange. In particular, the careful specification of conceptual domain information was critical for designers who wanted their data to conform to an exchange model. In contrast, users associated with MPI began to realize that they needed a repository made up of linguistic *concepts* used somewhat like thesaurus labels for flexible data retrieval rather than one containing formally defined DCs with rigorously controlled conceptual domains. There is, in fact, an important distinction to be made between a *linguistic concept* and a *linguistic DC*, a distinction that was not, and still has not been, explicitly articulated. The data model of the DCR was not ideal for documenting such concepts; it contained metadata and structures needed to describe DCs, but not needed, or even counterproductive, for describing linguistic concepts, and MPI users therefore found it unnecessarily complex. This, coupled with a shift in resource allocations at MPI, led MPI to decide to withdraw from the DCR at the end of 2014. While initially disconcerting to TC 37, MPI's decision led to an opportunity to review the current system and operational framework with an aim toward improving usability, quality, and integrity.

After evaluating potential replacement systems, TC 37 selected TermWeb, which is a terminology management system offered by Interverbum Technology. TermWeb is fully adaptable for managing discrete units of content of various sorts, not just terminology, and it turned out to be a suitable application for housing the content of the DCR. The primary community to be served by the DCR in its new configuration remains terminologists designing termbase models, particularly those working in the context of ISO 30042, as well as ISO 12616 for *Translation-oriented terminography.* The new DCR is hosted on a TermWeb database instance. A complementary website, datcatinfo.net, acts as a gateway to the TermWeb resource and provides information about the project. It is closely coordinated with tbxinfo.org, a web-based compendium of specifications and utilities designed to facilitate the creation, manipulation, and exchange of terminological data, especially in XLIFF-aware environments (ISO 21720). Implementation of TBX dialects designed for efficient and accurate exchange of termbase content is specifically linked to the DCs recorded in the DCR. A second community that relies on the DCR is supported by the Lexical Markup Framework (LMF) standard, ISO/WD 24613–5. Future plans for developing a mapping tool between TBX and LMF rely on the coordination of coherent data categories between the two standards.

In December 2014, the transition began from the web application developed by MPI to TermWeb. MPI switched the dynamic DCR it had developed to static (read-only) mode and provided the DCR Board (which had now devolved into a less-formal management committee) with copies of the static files containing the data category specifications. In parallel, the CLARIN group has maintained its Concept Registry since 2015.

Migrating the DCR to TermWeb

Migrating the data category specifications to TermWeb was not straightforward. Since the collection had grown under a crowd-sourcing model without coherent content management, the first task was to acquire a deep understanding of the data. Given the concerns over quality, it was essential to rank subsets of DC specifications according to confidence level, as well as to identify and remove the parts of the data model that were either redundant or no longer necessary.

The approach adopted was to study the schema for the Data Category Interchange Format (DCIF), which was the resident XML markup language for representing data category specifications within ISOcat, that is to say, in the original DCR. DCIF comprised 43 elements, 18 attributes, and 43 data types—a total of 104 different instances of markup strings. The DCR comprised 6,185 DC specifications, each one being documented in a separate file named <number>.dcif. For instance, /part of speech/ was 396.dcif.

Statistical analysis using the WordSmith Tools concordancer revealed how the 104 DCIF strings were actually used in the 6,185 dcif files. The advantage of WordSmith over other concordancers is that it allows batch searches, which meant that all 104 DCIF strings could be submitted for analysis across all 6,185 dcif files at once. (Other concordancers available at the time only allowed one string to be searched at a time.) WordSmith produces a batch report that shows the frequency of occurrence of each string. It was thus determined that six elements, five attributes, and nearly all the data types—20% of the total markup artifacts—were absent or so rare in the dcif files that they could be eliminated without loss of information.

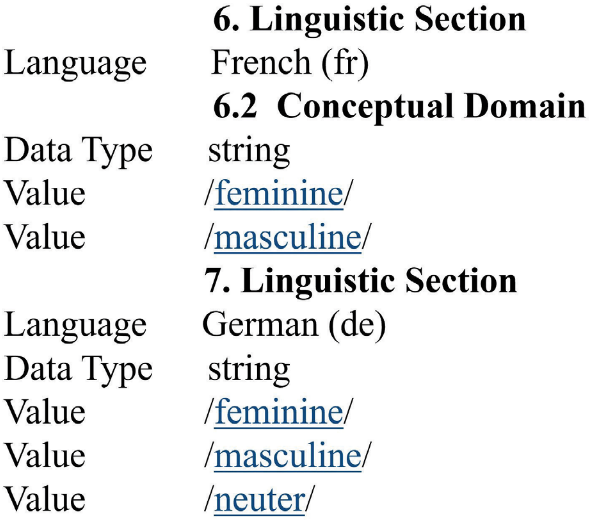
In the old DCR, the person who originally created a DC was recorded as its "owner." Initially, the DC was given a "private" scope value; the owner was the only person who could access it. When the owner felt that the DC could be of interest to a wider community and was comfortable sharing it, he or she could change the scope to "public," which made it available to other users. There were 1,954 private DCs and 4,231 public DCs in the set supplied by MPI.

WordSmith results also revealed frequent duplications, unusual or questionable content, incorrectly used fields, and other problems. These problems occurred less frequently in the public DCs compared to the private ones, the former benefitting from the checks and balances of the crowd. For this reason, the public DCs were migrated first.

As stated earlier, the Linguistic Section was intended to allow language-specific examples, explanations, and conceptual domains; it was intended to be a subset of the DC-level conceptual domain. A frequently cited case of the need for language-specific conceptual domains is the DC /grammatical gender/ (1297), where permissible values for French are /masculine/ and /feminine/, whereas German also allows /neuter/, as shown in figure 5.3.

However, the Linguistic Section had been the subject of negative user feedback: It constituted an additional nested level in the data model, added complexity to the user interface, and had proven difficult for users to apply correctly.

It turned out that only 37 DCs contained a Linguistic Section, or about 0.6%. Furthermore, in most cases, it had been misused. For instance, in nine DCs it was empty. In 22 DCs there was only one Linguistic Section, and it either did not provide any additional information or it contained only an example, which could easily be moved to the higher



	6. Linguistic Section
Language	French (fr)
	6.2 Conceptual Domain
Data Type	string
Value	/feminine/
Value	/masculine/
	7. Linguistic Section
Language	German (de)
Data Type	string
Value	/feminine/
Value	/masculine/
Value	/neuter/

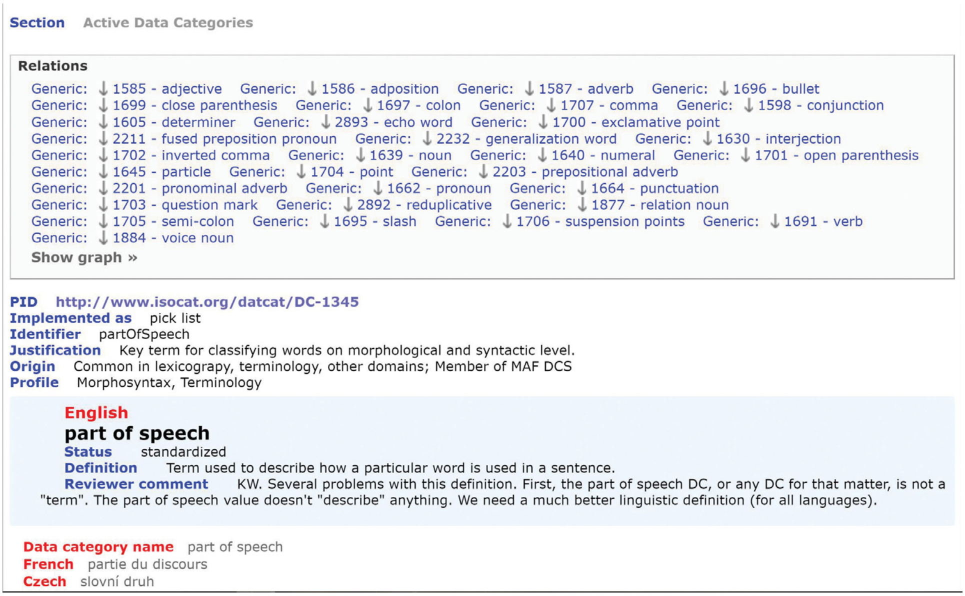
Figure 5.3
Two linguistic sections for /gender/.

Language Section. Only six public DCs (less than 0.2%) contained Linguistic Sections that could be justified.

The original intent of the Linguistic Section remains valid, as it is sometimes necessary to say something about a DC for a specific language. However, creating a distinct structure in the data model for such a rarely occurring feature is not justified. Such information can be recorded in a Note or other field of the Language Section. Eliminating the Linguistic Section simplifies the data model considerably, makes the new DCR (the Data Category Repository) easier to use, and eliminates data redundancy.

Another key finding relates to the conceptual domains themselves. Some DCs have, or were envisioned to have, different conceptual domains for different linguistic applications. For example, the permissible values of a DC could be different for terminology resources as opposed to another type of language resource, such as morphological annotation schemes. However, again it was interesting to consider whether the incidence of different conceptual domains for different application areas was statistically significant. Confining DCs to one conceptual domain at a time would further simplify the new data model.

The analysis indicated that only four public DCs (fewer than 0.1%) had more than one valid conceptual domain. This statistical evidence cast doubt on the need for allowing multiple conceptual domains in a DC. For the rare cases where application-specific conceptual domains are needed, such as the /part of speech/, which requires dozens of values for Morphosyntax but only a few for other applications, creating separate DCs for each case would



Section Active Data Categories

Relations

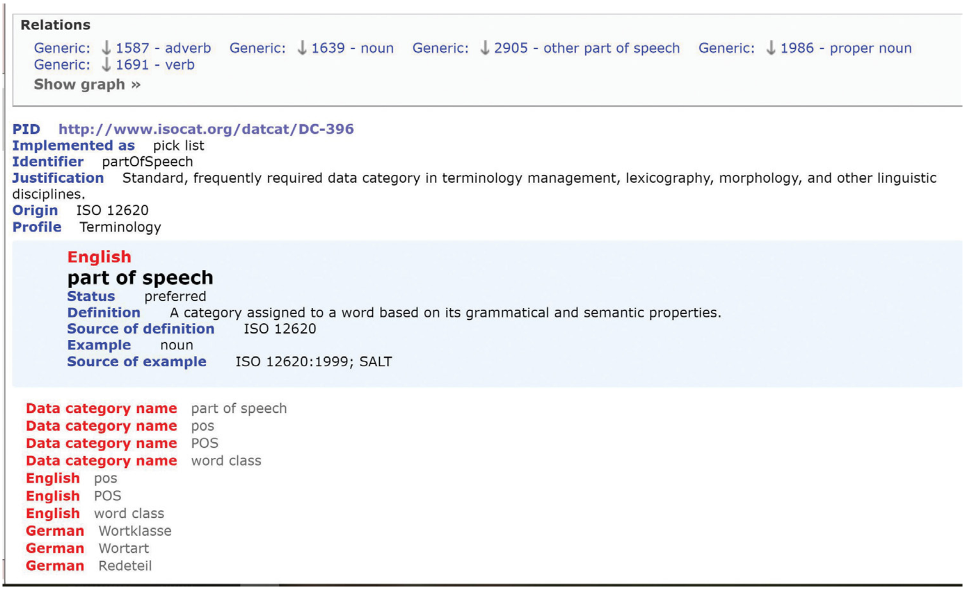
Generic: ↓ 1585 - adjective Generic: ↓ 1586 - adposition Generic: ↓ 1587 - adverb Generic: ↓ 1696 - bullet
Generic: ↓ 1699 - close parenthesis Generic: ↓ 1697 - colon Generic: ↓ 1707 - comma Generic: ↓ 1598 - conjunction
Generic: ↓ 1605 - determiner Generic: ↓ 2893 - echo word Generic: ↓ 1700 - exclamative point
Generic: ↓ 2211 - fused preposition pronoun Generic: ↓ 2232 - generalization word Generic: ↓ 1630 - interjection
Generic: ↓ 1702 - inverted comma Generic: ↓ 1639 - noun Generic: ↓ 1640 - numeral Generic: ↓ 1701 - open parenthesis
Generic: ↓ 1645 - particle Generic: ↓ 1704 - point Generic: ↓ 2203 - prepositional adverb
Generic: ↓ 2201 - pronominal adverb Generic: ↓ 1662 - pronoun Generic: ↓ 1664 - punctuation
Generic: ↓ 1703 - question mark Generic: ↓ 2892 - reduplicative Generic: ↓ 1877 - relation noun
Generic: ↓ 1705 - semi-colon Generic: ↓ 1695 - slash Generic: ↓ 1706 - suspension points Generic: ↓ 1691 - verb
Generic: ↓ 1884 - voice noun
Show graph »

PID http://www.isocat.org/datcat/DC-1345
Implemented as pick list
Identifier partOfSpeech
Justification Key term for classifying words on morphological and syntactic level.
Origin Common in lexicograpy, terminology, other domains; Member of MAF DCS
Profile Morphosyntax, Terminology

English
part of speech
Status standardized
Definition Term used to describe how a particular word is used in a sentence.
Reviewer comment KW. Several problems with this definition. First, the part of speech DC, or any DC for that matter, is not a "term". The part of speech value doesn't "describe" anything. We need a much better linguistic definition (for all languages).

Data category name part of speech
French partie du discours
Czech slovní druh

Figure 5.4
The /part of speech/ DC for Morphosyntax.



Relations

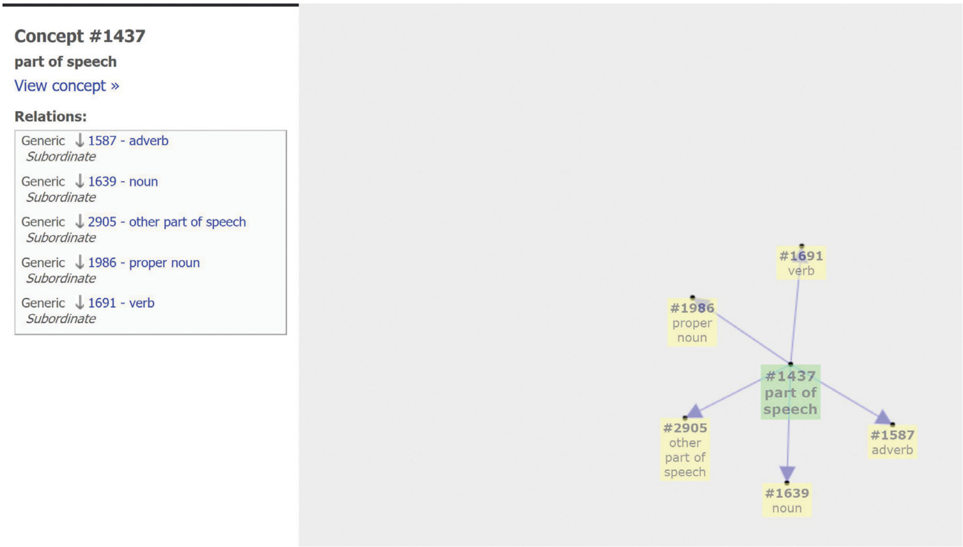
Generic: ↓ 1587 - adverb Generic: ↓ 1639 - noun Generic: ↓ 2905 - other part of speech Generic: ↓ 1986 - proper noun
Generic: ↓ 1691 - verb
Show graph »

PID http://www.isocat.org/datcat/DC-396
Implemented as pick list
Identifier partOfSpeech
Justification Standard, frequently required data category in terminology management, lexicography, morphology, and other linguistic disciplines.
Origin ISO 12620
Profile Terminology

English
part of speech
Status preferred
Definition A category assigned to a word based on its grammatical and semantic properties.
Source of definition ISO 12620
Example noun
Source of example ISO 12620:1999; SALT

Data category name part of speech
Data category name pos
Data category name POS
Data category name word class
English pos
English POS
English word class
German Wortklasse
German Wortart
German Redeteil

Figure 5.5
The /part of speech/ DC for Terminology.



Concept #1437

part of speech

View concept »

Relations:

Generic ↓ 1587 - adverb
Subordinate

Generic ↓ 1639 - noun
Subordinate

Generic ↓ 2905 - other part of speech
Subordinate

Generic ↓ 1986 - proper noun
Subordinate

Generic ↓ 1691 - verb
Subordinate

Figure 5.6
Relation diagram showing the conceptual domain for /part of speech/ for the Terminology profile.

be a reasonable solution. It could be argued that when the permissible values of a DC diverge considerably in different linguistic applications, what we are really dealing with is different data category concepts. This was, indeed, the perspective already taken by users of the DCR; it contained seven different DCs covering the concept of /part of speech/. It was therefore decided to disallow multiple application-specific conceptual domains for a DC.

Figures 5.4 and 5.5 show two DC specifications for /part of speech/ from TermWeb. Figure 5.4 is the DC configured for Morphosyntax, and figure 5.5 is the DC applied to Terminology. Note the differences in the conceptual domain, shown as Relations in TermWeb. The conceptual domain of /part of speech/ for Morphosyntax comprises many more members than that for Terminology. Figure 5.6 shows the members of the conceptual domain for Terminology in a diagram format.

With the participation of a global community of stakeholders, the DCR had become a crowd-sourced resource, operating for all practical purposes outside the formal ISO environment. The workflows that were developed to permit standardization of DCs according to the traditional ISO stages were never used. Indeed, during six years of operation, not a single DC specification was standardized in the DCR. This fact could not be ignored and led the management committee to recognize that the DCR served a harmonization, rather than a standardization, role. For this reason, the standardization workflow sections of the DCIF model, which included major parts of the Administration Section, were eliminated.

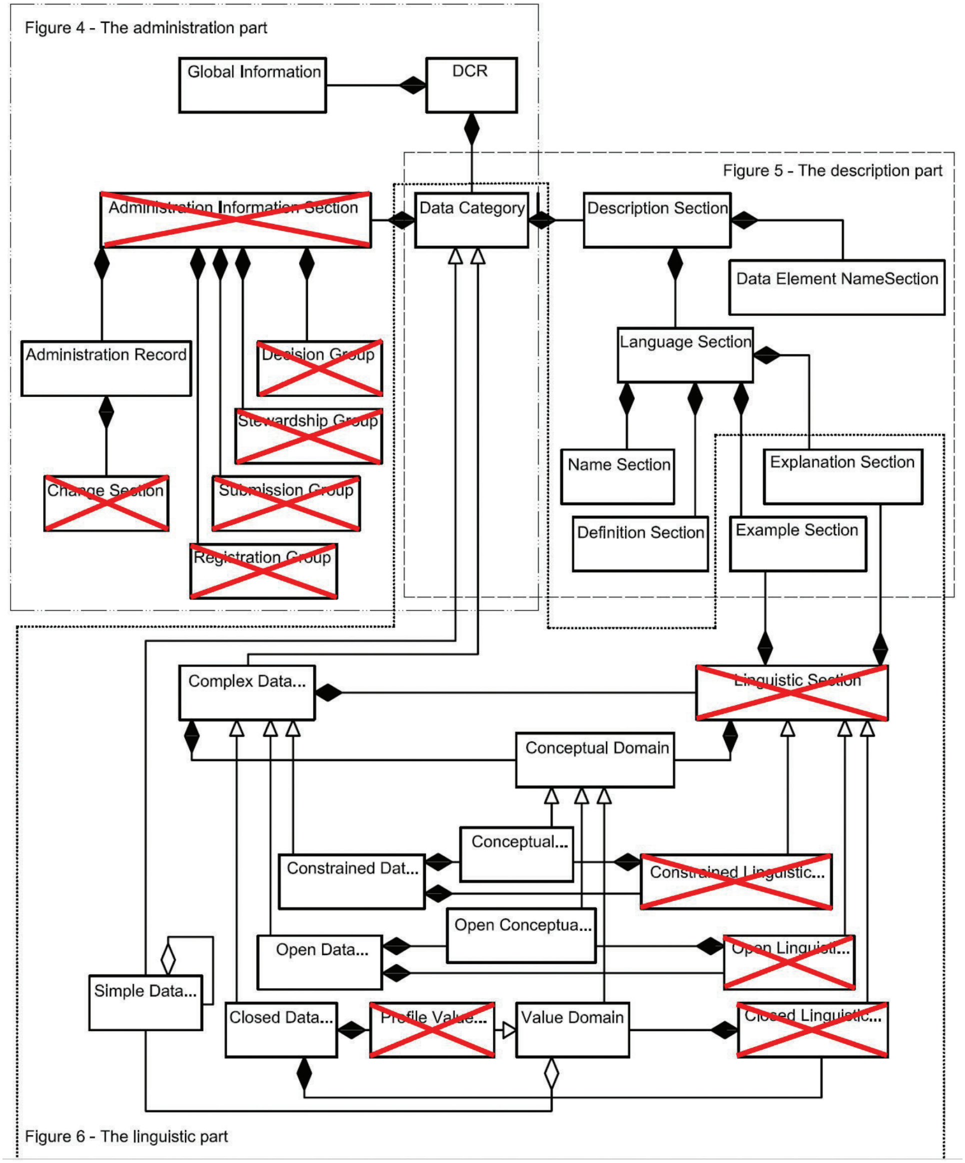


Figure 5.7
Original Figures 4–6 from ISO 12620:2009: Data model for the new DCR showing parts removed from ISOcat

The DCR supports documenting DC specifications in 37 languages, and more can be added. However, currently, there is no content for 11 languages, and very little for a handful of others. At the same time, questions have been raised about the need for any languages other than English. Indeed, the mandate of the DCR is to describe DCs, not to "translate" them. On the other hand, the TermWeb system supports multilingual content out-of-the-box, so the multilingual nature of the resource has been maintained, even though doing so meant that it is necessary to review and maintain the multilingual information alongside the English content.

To summarize, the Linguistic Section and the parts of the Administrative Section that supported the ISO standardization workflow have been eliminated, and the Conceptual Domain has been restricted to only one instance per DC. Figure 5.7 shows the original data model with the removed components crossed out.

Converting DCIF to TBX

As stated earlier, the source files from the DCR were serialized in a specialized format called DCIF. TermWeb only supports import of XML files that are in TermBase eXchange (TBX) format. Therefore, the DCIF files needed to be converted to TBX, which is sufficiently granular to represent the DCIF components that were retained after analysis and revision. However, DCIF represents data categories, while TBX represents terminology. Although the structure of the two data models was similar, there were sufficient differences to make the conversion quite challenging.

First, the DCIF elements and attributes were mapped to equivalent TBX structures. This involved not just changing the name of an element or an attribute, but sometimes also moving an element to a different location in the entry model, or converting information from an element name to an attribute value or from an attribute value to element content.

In June 2016, Kara Warburton, a co-author of this chapter, prepared a specification document outlining the mapping requirements. LTAC Global, a nonprofit consortium that supports initiatives promoting interoperability of language resources, generously provided the services of a software engineer who developed a conversion script based on the specification. The following rules were implemented in the script.

a) Remove unwanted DCIF markup

Markup representing information that would become redundant or irrelevant in the target system, such as historical dates, user names, and nesting elements that do not have TBX counterparts, needed to be removed.

Furthermore, the members of the conceptual domain of a closed DC (such as /noun/ for /part of speech/) could not be directly imported into TermWeb. The individual simple DCs themselves (/noun/, etc.) were imported automatically by the migration script, but their membership in the parent closed DC could not be represented in the import file. This is because the link between a simple DC and its parent is established in TermWeb via a "relation," and relations are not importable. Those elements were therefore also removed from the parent DC specifications in the import file. After import, the relations were established manually by linguists working in TermWeb. Table 5.1 provides examples of markup removed from the original DCIF.

Table 5.1
Markup deleted from DCIF

Reason for removal	Example
Unnecessary nesting element	<dcif:definitionSection>
Unwanted details	<dcif:effectiveDate>
Statistically irrelevant part of data model	<dcif:linguisticSection type="closed">
Members of a closed conceptual domain (to be added later manually)	<dcif:conceptualDomain type="closed"> <dcif:profile>Terminology</dcif:profile> <dcif:value pid="http:// … "/> … </dcif:conceptualDomain>

b) Convert DCIF markup to TBX

The script converted DCIF markup to TBX markup. Tables 5.2–5.4 show some of the main types of conversions.

Table 5.2
Mapping between DCIF and TBX elements

Straightforward mapping	
DCIF	**TBX**
<dcif:justification>	<descrip type="justification">
<dcif:identifier>	<descrip type="identifier">
<dcif:name>	<term>

Table 5.3
Element mergers

Two elements becoming one	
DCIF	**TBX**
<dcif:languageSection> <dcif:language>fr</dcif:language>	<langSet xml:lang="fr">

Table 5.4
TBX markup variations

Additional precision and conversions in the TBX rendering	
DCIF	**TBX**
<dcif:source> (in a definition section)	<descrip type=”sourceOfDefinition”>
<dcif:dataElementName>	<langSet xml:lang=”eo”> <tig> <term>
<dcif:dataCategory pid=”http:// … “type=”simple”>	<descrip type=”PID”>http:// … </descrip> <xref type=”externalCrossReference” target=“http:// …”>http:// … </xref> <descrip type=”dataCategoryType”>value</descrip>

The script was used to convert the DCIF files in nine batches of about 500 files each. Only the 4,327 public DCs were submitted to this process; the 1,962 private ones were retained in an archive. Each converted batch resulted in a single merged TBX file.

Reviewing the TBX Files

The next step was to manually review each of the nine files before import. During this review, a few problems in the conversion process were found and reported to the software engineer, who updated the conversion script, and the conversion was then repeated. These problems were in most cases attributed to missing or incorrect information in the migration specification, usually because the full range of information types and instances in thousands of dcif files could not be anticipated.

The review made it possible to identify and address many content-related problems before importing the DC specifications into TermWeb. The changes that were made include the following:

- Eliminating redundancy: For instance, the same bibliographical reference was often cited repeatedly in the same DC, and occasionally multiple fields contained duplicate information (such as Justification and Definition).
- Moving information to the correct places: For instance, if a Definition was actually an Explanation or a Note, it was moved accordingly.
- Splitting combined information to separate fields: For instance, when a Definition field included both a definition and a note, the note part was moved accordingly.
- Resolving incomprehensible or cryptic notations: For instance, acronyms used for people’s names and abbreviated forms of various types. Potentially unfamiliar abbreviations, such as T9n/L10n, were expanded to their proper formulations.

- Removing obsolete or meaningless notations, such as “green text,” which was a historical notation no longer relevant.
- Fixing typographical errors and spelling mistakes.
- Fixing formatting problems.
- Removing elements that contained only placeholder text.
- Removing Origin values that are too general, for instance “linguistics literature.”
- Checking URLs and removing or replacing broken links.
- Fixing a number of incorrect DC names.

Other more-substantive changes were recommended, but it was felt that substantive changes should be discussed with representative stakeholders beforehand. For this purpose, a field called “Reviewer comments” was created in TermWeb, along with a corresponding element in the TBX import file. A total of 385 reviewer comments were included in the import file. Anyone working in DatCatInfo can use this field to record suggestions.

To preserve a copy of the original data, should any of the changes be questioned in the future, in most cases XML commenting tags were inserted around the original content and the new corrected version was added alongside the original. There are now 4,984 sets of commenting tags.

Considering the number of reviewer comments and XML commenting tags, nearly 5,400 edits, changes, and suggestions were made to the imported DC specifications. While more work is still needed, significant progress has already been made in addressing previous concerns about quality.

Another problem was duplicate entries.

Duplicates

During the review, a significant number of duplicate DCs, or DCs that are potentially duplicate, were discovered. Various text analysis tools were used to measure the scope of duplication, based on comparing the DC names. Five percent (160) of the imported DCs have the same name as another DC. These DCs will have to be checked to determine which are actual duplicates, and their DC specifications harmonized. For the DCs not yet imported, the potential duplication is larger: 10% (335). It will be necessary to address those duplicates before import.

These figures represent DCs whose names are identical. But duplicate DC specifications also occur where the names are not identical. Some offer clues based on similarities in the DC name—for instance, /judicial interpreting/, /judiciary interpretation/, and /judiciary interpreting/. Others have no similarities in the names whatsoever, yet examination of other metadata can confirm that in fact they refer to the same DC concept. This issue represents a major aspect of the future harmonization activities: The entire DCR needs to be reviewed to identify and resolve duplicates.

Aside from duplicates, there are also cases where the DC's status as a DC was questionable.

What Is a Data Category, and What Is Not?

For some DC specifications, doubts were raised whether what was being described was a data category. The first group of questionable DC specifications comprises the "linguistic concepts" entered to meet needs in the CLARIN project. A linguistic concept is not necessarily also an instance of actual data in any language resource. For instance, DC 3998 is /language for special purposes/ (LSP). The DC definition was taken from ISO 1087, which is a glossary, and is already publicly available in the ISO OBP. Is "language for special purposes" also a data category used in some language resource? Quite possibly. The danger is in accepting without question linguistic concepts into the DCR. A clear distinction needs to be made between linguistic concepts and linguistic data categories, with pure concepts probably being isolated in an archive.

The second group of questionable DC specifications includes code strings, such as the name of an element or an attribute from an XML vocabulary or annotation scheme. Many of this type originated from the Text Encoding Initiative (TEI). For example, DC 6186, simply called /a/, is from the TEI header. Another example is DC 2794, /ADJA/, which is described as the "STTS tag for attributive adjective." But the DCR already has two DCs called /attributive adjective/ (1243 and 5242). Most likely, then, /ADJA/ is merely a code representation of one of these existing DCs, where it should be added as an alternative DC name. How code representations should be handled needs to be decided.

The third group covers DCs from nonlinguistic domains, such as medical/scientific concepts. For example, DC 4458 describes /magnetoencephalography/, and includes a description from Wikipedia. Should the DCR even include such information? These DCs probably represent data points used in documenting language-related physiological testing, but they aren't normally associated with language resources per se.

DCs from external sources, such as Edisyn (2011), the GOLD ontology (2010), and STTS (1995/1999), are already documented and maintained by their source organizations. Whether or not to include DCs that are already documented in another public resource is another debatable question. On the one hand, doing so represents a form of duplication and redundancy, and it would be virtually impossible to ensure that the DC specification is always up-to-date and synchronized with its source version. On the other hand, one purpose of the DCR is to offer a trusted source of information about DCs in one convenient location. Having information about DCs in one location fosters harmonization. Rejecting all DCs that are documented in an existing public resource would run contrary to the objectives of the DCR. An official decision in this regard has not been made. In the meantime, the DCs from these three organizations have been temporarily excluded. One viable solution would be to maintain a cross-reference entry in the DCR that would point users to other comparable concept, term, or label registries.

As with any database, there should be clear criteria for what qualifies for inclusion. Unfortunately, there appear to be no documented inclusion criteria for the DCR. The new ISO 30042 and 12620 offer a definition for *data category* (ISO 30042, 3.8):

data category

class of data items that are closely related from a formal or semantic point of view

EXAMPLE: /part of speech/, /subject field/, /definition/

Note 1 to entry: A data category can be viewed as a generalization of the notion of a field in a database.

Nevertheless, the range of content described in the DC specifications that were reviewed before migration suggests that the many contributors to the collection were operating without any consensus of what a data category actually is. In the absence of clear guidelines, in many cases the questionable DC was kept in the import file and a reviewer comment was included to draw attention to the issue. Types of DCs that were removed from the import files, and kept in an archive for future consideration, are shown in table 5.5.

Table 5.5
Data categories removed from import files

Number of DCs	Issue or concern
5	Complex DCs that do not have a conceptual domain
13	DCs that are related to the DCR itself
469	Deprecated language codes
12	Odd or strange DCs[2]
2	Linguistically incorrect DCs
1	Deprecated DCs
5	Rendering problems
1	Unknown source
19	Constrained
73	Container type
92	From Edisyn
494	From the GOLD Ontology
54	From STTS
8	Superseded
1,248	TOTAL

After excluding the DCs shown in table 5.5 and the private ones previously mentioned, the remaining 2,975 DCs were imported into TermWeb in September 2016. Future tasks for the DCR team will include the resolution of open questions from the above discussion.

Post-Import Work

As already described, the first major task after import was to establish the links between DCs with closed conceptual domains and simple DCs that are the values of those domains. For instance, it was necessary to link /part of speech/ (DC 396) with /noun/ (DC 1639), /verb/ (DC 1691).

Other pending tasks include addressing all the reviewer comments and deciding whether to eventually import the DCs that were held back in an archive during this initial import. There is also the question of ongoing maintenance, additional cleaning, vetting, harmonization of duplicates, and addition of new DCs.

A Shifting Identity and Purpose

During the migration between 2014 and late 2016, the committee in charge operated as an extension of ISO/TC 37/SC 3/WG 1 (Data Categories), and its regular progress reports were presented at the ISO/TC 37 annual meetings. It became apparent that the original purpose and mandate of the DCR, as defined in ISO 12620:2009, had changed, or needed to change, since the standardization mission was never fulfilled or even desired. At most, users needed a trusted source of information about DCs. This can be achieved through a less-formal process of consolidation, review, and harmonization.

Furthermore, the DCR had suffered "scope-creep" in the application areas, which are referred to as thematic domains, that is, the disciplines covered within the broad category of linguistics. ISO 12620:2009 stated that the DCR is "applicable to all types of language resources" without offering a definition of what qualifies as a language resource for this purpose. However, in several other places in the standard, it is stated that the intent of the DCR was to cover data categories required for ISO/TC 37 standards alone, as the following quotations demonstrate:

> It shall provide a reference repository for data categories and related information for all the existing or future standards in ISO/TC 37 that involve data modelling or data interchange (Clause 5).

> The creation of a single global Data Category Registry (DCR) for all types of language resources treated within the ISO/TC 37 environment provides a unified view of the various applications of such a reference resource (Introduction).

> The DCR will eventually contain all ISO/TC 37 data categories.… (Introduction).

Also in the Introduction, four thematic domains "have been recognized as definable subsets of the DCR": Terminology, Semantic Content Representation, Language Codes, and Lexicography. Yet Language Codes are already maintained and made publicly available by the US Library of Congress (LoC, 2013) and Ethnologue (2015). The related Country Codes are available through the ISO Online Browsing Platform (ISO 2018). Two thematic domains not mentioned in ISO 12620:2009 grew disproportionately large in the DCR,

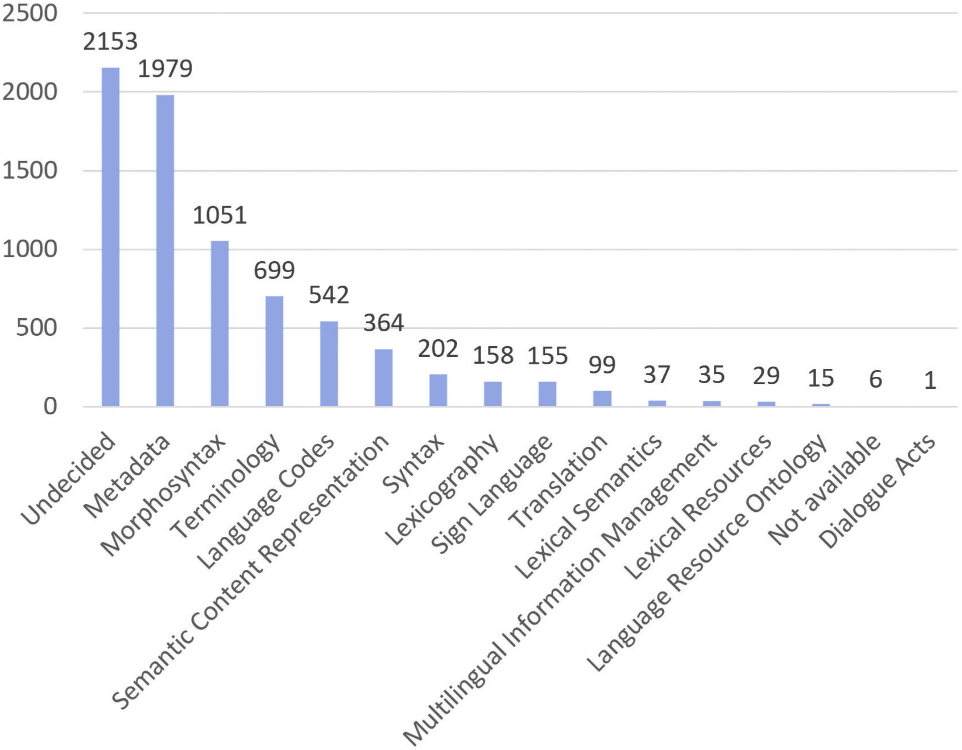


Figure 5.8
Number of DCs assigned to each thematic domain.

while the four explicitly mentioned remained underdocumented. The thematic domain most frequently used was "Undecided." Other DC specifications, as already mentioned, described concepts that are not part of any linguistic domain. Figure 5.8 shows the number of DCs that were assigned to each thematic domain.

The Introduction in ISO 12620:2009 also states that "it is not the intent of this International Standard to define an ontology of language resources." In retrospect, this was an unfortunate omission, as an ontology of language resources would have improved the use of the thematic domains, which were meant to categorize DCs into logical subsets. (The reason for this decision rested in an earlier attempt to classify the DCs in ISO 12620:1999, an effort that was widely viewed as a failure, primarily because the multifaceted nature of the DCs tends to preclude any mono-layered classification.) Consequently, in the DCR, the use of thematic domains was inconsistent, there was significant overlap between the thematic domains themselves, the value "Undecided" was extensively used, and DCs were frequently assigned to multiple thematic domains simultaneously. The latter is expected since DCs are often used in various types of language resources. Nevertheless, it was clear that the assignment of multiple thematic domains to a DC was not always justified. The use of thematic domains in general was highly problematic.

Given the shift away from a formal standardization role and the difficulties associated with the thematic domains, the managing committee decided to revisit the mandate and scope of the DCR. Formal standardization was abandoned in favor of less-formal harmonization and, in view of the quality issues cited above and the confusion over thematic domains, it was decided to focus immediately on the original thematic domain—Terminology—in order to prioritize the cleanup work. Thus, the DCs relating to terminology resources will be reviewed first in the new TermWeb environment.

Rebranding: DatCatInfo

The shift from standardization to harmonization as a purpose meant that the new DCR was, in effect, no longer an ISO resource. Long discussions about the ramifications of these changes ensued between the managing committee and ISO Central Secretariat. It was mutually agreed that the DCR was not—and never had actually been—a data category *registry*, but rather had always served the less-formal purpose of a *repository*. This is why it was agreed that the acronym DCR would henceforth represent *Data Category Repository*.

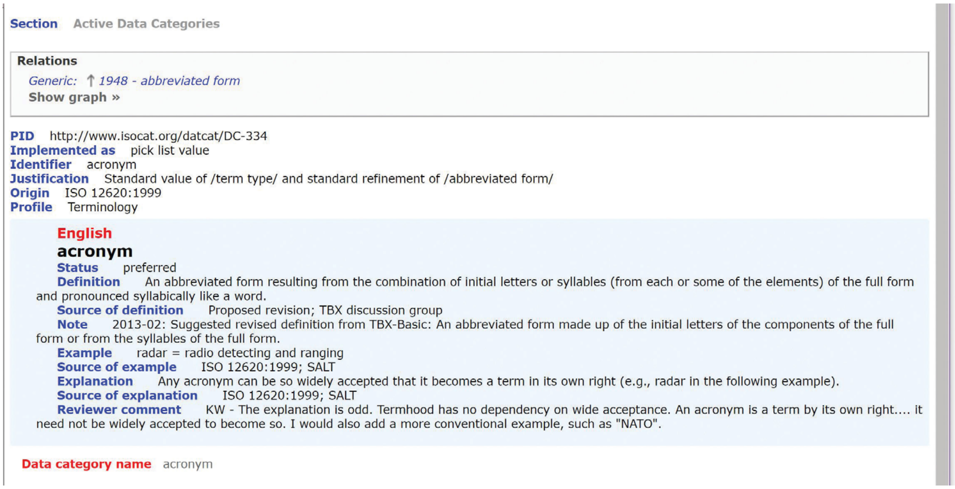
Since the DCR was no longer viewed as an ISO resource, using the existing brand name *ISOcat* and the URL *www.isocat.org* was also no longer permitted. The managing committee concluded an agreement with ISO Central Secretariat recognizing LTAC/TerminOrgs[3] as the owner of the DCR. The name of the web domain was changed to *DatCatInfo* and the URL to *www.datcatinfo.net*. Nevertheless, a search for www.isocat.org will redirect to www.datcatinfo.net, and the DCR is closely linked to www.tbxinfo.net, which provides information and utilities for the TBX standard.

As a consequence of these changes, ISO 12620:2009 was withdrawn and a new version has been produced for publication in 2019. It describes best practices for developing a Data Category Repository generically. As a consequence, the DCR itself is no longer a normative resource owned by ISO; rather, it is a collection of industry-harmonized, industry-sanctioned DCs.

Example from DatCatInfo

Figure 5.9 shows a data category specification in DatCatInfo. Here are a few observations worth noting:

- The Relation, showing that this DC has a generic relation pointing up to DC 1948 (abbreviated form). This means that /abbreviation/ is a simple DC and a member of the conceptual domain of /abbreviated form/. The DC type *simple* can also be determined by the *Implemented as* field, which indicates *pick list value*.



Section Active Data Categories

Relations
Generic: ↑ *1948 - abbreviated form*
Show graph »

PID http://www.isocat.org/datcat/DC-334
Implemented as pick list value
Identifier acronym
Justification Standard value of /term type/ and standard refinement of /abbreviated form/
Origin ISO 12620:1999
Profile Terminology

English
acronym
Status preferred
Definition An abbreviated form resulting from the combination of initial letters or syllables (from each or some of the elements) of the full form and pronounced syllabically like a word.
Source of definition Proposed revision; TBX discussion group
Note 2013-02: Suggested revised definition from TBX-Basic: An abbreviated form made up of the initial letters of the components of the full form or from the syllables of the full form.
Example radar = radio detecting and ranging
Source of example ISO 12620:1999; SALT
Explanation Any acronym can be so widely accepted that it becomes a term in its own right (e.g., radar in the following example).
Source of explanation ISO 12620:1999; SALT
Reviewer comment KW - The explanation is odd. Termhood has no dependency on wide acceptance. An acronym is a term by its own right.... it need not be widely accepted to become so. I would also add a more conventional example, such as "NATO".

Data category name acronym

Figure 5.9
The DC /acronym/ in DatCatInfo.

- The Persistent Identifier (PID), which reflects the file name (334) of the original DC specification from the former Data Category Registry. The isocat.org domain name in the PIDs will eventually be updated to datcatinfo.net.
- The Identifier, "acronym." This is the machine-readable name of the DC. In compound names, the identifier is therefore written in camel case, for instance, *partOfSpeech* for /part of speech/.
- The Profile, "Terminology," also known as the thematic domain.
- The English name, "acronym," and the Data category name, "acronym" (red fields in the e-pub version). The latter is meant to be a language-agnostic human readable name of the DC. The *Data category name* field is not filled in for all DCs because some users did not realize its importance, so a number of DCs only have an *English name*. For this reason, when searching in DatCatInfo, it is best to choose *English* as the search language.
- The Reviewer comment, to be addressed during revision.

Current Status, Future Work, and Challenges

This chapter has described the migration of language resource data categories from the original Data Category Registry to a new Data Category Repository, whereby the intact data collection is referenced jointly by the abbreviation DCR. Approximately half the DCs from the Data Category Registry have been migrated to the new DCR, called DatCatInfo in the TermWeb environment (totaling 2,977 DCs). The remaining DCs have been kept in a static archive.[4] However, much work remains to address the reviewer comments, harmonize duplicates, reconsider the archived DCs, and complete other cleanup tasks. The primary challenge in this endeavor will be coordinating the work and addressing it in stages. Thanks to the efforts of a team of volunteers, DatCatInfo has emerged as a new and improved free public resource from a former ISO project that could have otherwise been canceled entirely. Given the scope of work and support that it requires, financial support is urgently needed to achieve its stated goals. The managing committee is searching for grant opportunities.

Some challenges are purely technical. DC specifications contain a Persistent Identifier (PID), for example: http://www.isocat.org/datcat/DC-1840. With the transfer to DatCatInfo, all PIDs are being changed accordingly; for example, http://www.datcatinfo.net/datcat/DC-1840. The work of converting the PIDs is still in progress. Old ISOcat PIDs embedded in legacy resources will resolve to the new environment, thus maintaining the requirement for persistence inherent in the system.

The governance procedures outlined in ISO 12620:2009 do not apply to DatCatInfo. As noted above, a new version of ISO 12620, describing the management of a DCR generically, has been approved by ISO/TC 37 for a 2019 publication date. While DCs are no longer intended for formal standardization, procedures are still needed for harmonizing duplicates, improving content, deprecating DCs, and accepting new ones. New DCs will need to come from end-users, so a contribution workflow will need to be implemented. (Currently TermWeb has a feature for submitting feedback, but it will not suffice.) Defining these fundamental components—governance procedures and public contribution workflows—is one of the most pressing tasks for the managing committee.

The lack of inclusion criteria is a serious shortcoming. Such criteria need to be determined. Key questions include:

1. What is a language resource? What types of language resources are served by the DCR?
2. What is a data category? What is not a data category?
3. How does a data category differ from a linguistic concept?
4. What are the linguistic domains that the DCR should cover? Can they be clearly defined?
5. How can we clearly determine what linguistic domain a DC applies to?
6. What criteria should be used to determine whether DCs already available in a public resource should also be included in the DCR? How should they be included so as to avoid redundancy and maintain currency?

The greatest challenge in developing and maintaining DatCatInfo is the lack of funding. For instance, developing the required contribution workflow will require programming resources. All the work is currently being carried out by volunteers on an ad hoc basis.

This chapter has traversed the history of the DCR, following it from a purely paper standard, through its history as a Data Category Registry intended for administration by an ISO Registration Authority, to a Data Category Repository freely available on the web under a Creative Commons license. The committee responsible for the collection will continue to address the harmonization and issues described in this chapter but, as noted, there is no clear timeline for completing this work.

What can be learned from this experience in order to avoid the variation in mission and discordant goals that have marked the evolution of the DCR? Certainly, the lack of clear consensus on fundamental aspects, such as inclusion criteria and thematic domains, cannot be attributed to any negligence on the part of ISO/TC 37/SC 3, which set up and operated a DCR Governance Board for years and held countless meetings and consultations in an effort to define and achieve common goals. Extensive discussion, planning, collaboration, and goodwill were invested in the project, and even when faced with difficult events, such as the loss of MPI as a major contributor, the work was amicable and devoid of any misunderstandings or conflicts. Indeed, the divergent needs and goals reflect a paradigm discontinuity between the various communities of practice that came together in good faith. As Thomas Samuel Kuhn warns, the confident mastery of essentially the same terminology used to define the project in the end masked divergent needs and practices. Perhaps a more targeted analysis from the outset might have revealed issues of this nature earlier, but again, if that noted philosopher of science is our guide, these kinds of indeterminacies are inevitable.

Notes

1. Data category names when cited in running text are enclosed in forward slashes (e.g. /part of speech/).
2. Examples of the "odd or strange" group include DCs that were obviously created for testing purposes, such as DC 1500, /coca cola/, and several DCs that had no description whatsoever.
3. LTAC/TerminOrgs is a liaison organization to ISO/TC 37.
4. http://www.datcatinfo.net/rest/user/guest/workspace.

References

CLARIN. 2015. CLARIN Concept Registry (CCR). Accessed January 2, 2018. https://concepts.clarin.eu/ccr/browser/.

CLARIN. 2017. CLARIN: European Research Infrastructure for Language Resources and Technology. Accessed January 2, 2018. https://www.clarin.eu/.

DatCatInfo. Accessed September 15, 2018. www.datcatinfo.net.

Edisyn. 2011. European Dialect Syntax Search Engine. Accessed January 2, 2018. http://www.meertens.knaw.nl/edisyn/searchengine/.

ELRA/ELDA. 2017. What Is a Language Resource? Accessed January 2, 2018. http://www.elra.info/en/about/what-language-resource/.

Ethnologue. 2015. ISO 639 Code Tables. Accessed May 31, 2019. https://www.ethnologue.com/.

Gold Community. 2010.. GOLD. Accessed January 2, 2018. http://linguistics-ontology.org/.

Interverbum Tech. 2018. TermWeb. Accessed September 9, 2018. http://demo.termweb.se/termweb/app.

ISO 704:2009. Terminology work—Principles and methods. Geneva: International Organization for Standardization.

ISO 8879:1986. Information processing—Text and office systems—Standard Generalized Markup Language (SGML).

ISO 12200:1999. Computer Applications—Machine-Readable Terminology Interchange Format (MARTIF). Geneva: International Organization for Standardization. Withdrawn.

ISO/CD:2018 12616–1. Terminology work in support of multilingual communication—Part 1: Fundamentals of translation-oriented terminography. Geneva: International Organization for Standardization. Under development.

ISO 12620:1999. Computer applications in terminology—Data categories. Geneva: International Organization for Standardization. Withdrawn.

ISO 12620:2009. Systems to manage terminology, knowledge and content—Specification of data categories and management of a Data Category Registry for language resources. Geneva: International Organization for Standardization. Withdrawn.

ISO 12620:2019. Management of terminology resources—Data category specifications. Geneva: International Organization for Standardization.

ISO 16642:2003. Management of terminology resources—Terminological Markup Framework (TMF). Geneva: International Organization for Standardization.

ISO 21720:2017 (OASIS). XLIFF (XML Localisation Interchange File Format). Geneva: International Organization for Standardization.

ISO/WD 24613–5 Language resource management—Lexical markup framework (LMF)—Part 5: LBX serialisation.

ISO 30042:2008. Systems to manage terminology, knowledge and content—TermBase eXchange (TBX). Geneva: International Organization for Standardization.

ISO 30042:2019. Management of terminology resources—TermBase eXchange (TBX). Geneva: International Organization for Standardization

ISO. 2009. Launching of The ISO Concept Database (ISO/CDB) will benefit standards users and developers. Accessed January 2, 2018. https://www.iso.org/news/2009/11/Ref1261.html.

ISO. 2018. Online Browsing Platform (OBP). Accessed January 2, 2018. https://www.iso.org/obp/ui/.

ISOcat. 2014. The ISOcat Data Category Registry. Withdrawn. Accessed August 1, 2018. https://www.iso.org/sites/dcr-redirect/dcr.html.

ISO/CDB. 2009. ISO Concept Database. Accessed September 10, 2018. https://www.iso.org/news/2009/11/Ref1261.html.

Kemps-Snijders, M., M. Windhouwer, P. Wittenburg, and S. E. Wright. 2009. "ISOcat: Remodeling Metadata for Language Resources." *Special Issue: Open Forum on Metadata Registries of the International Journal of Metadata, Semantics and Ontologies* (IJMSO) 4 (4): 261–276.

LoC. 2013. Library of Congress Codes for the Representation of Names of Languages: Codes arranged alphabetically by alpha-3/ISO 639–2 Code. Accessed January 2, 2018. https://www.loc.gov/standards/iso639-2/php/code_list.php.

LTAC Global. Language Terminology/Translation and Acquisition Consortium. Accessed January 2, 2018. http://www.ltacglobal.org/.

LTAC/TerminOrgs. 2017. DatCatInfo Data Category Repository (successor to the isocat.org Data Category Registry). Accessed September 15, 2018. http://www.datcatinfo.net

SALT. 2000. XML representations of Lexicons and Terminologies (XLT)—Default XLT Format (DXLT). Accessed January 2, 2018. http://www.ttt.org/oscar/xlt/dxltspecs.html.

STTS. 1995/1999. Institut für Maschinelle Sprachverarbeitung Tag Table. Accessed January 2, 2018. http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html.

TBXInfo. 2018. TermBase eXchange. Accessed August 8, 2018. http://www.tbxinfo.net/.

TerminOrgs. 2017. Terminology for Large Organizations. Accessed September 15, 2018. http://www.terminorgs.net.

Wright, Sue Ellen, and Gerhard Budin. 1994. "Data Elements in Terminological Entries: An Empirical Study." *Terminology* I-1:41–59.

WordSmith Tools. Windows software for finding word patterns. Accessed September 15, 2018. http://www.lexically.net/wordsmith/.

6 Describing Research Data with CMDI—Challenges to Establish Contact with Linked Open Data

Thorsten Trippel and Claus Zinn

Introduction

The CLARIN (Common Language Resources and Technology Infrastructure) research infrastructure for the Social Sciences and the Humanities (SSH) offers researchers access to a wide range of language-related research data and tools. The Virtual Language Observatory, for instance, gives users uniform access to nearly a million resources and tools using faceted search on metadata, and by employing Federated Content Search users can perform full-text searches across distributed databases. Additionally, WebLicht supports users to process language resources with predefined and user-defined tool workflows, while the Language Resource Switchboard helps users to connect resources with tools that can process them. The infrastructure depends on a common metadata framework that makes it possible to describe all types of resources to a fine-grained level of detail, paying attention to their specific characteristics and the needs of the many SSH communities.

The Component MetaData Infrastructure (CMDI) follows a Lego brick approach to metadata modeling, where elementary data descriptors are semantically grounded in concept registries, and where components can be defined in terms of those descriptors or predefined simpler components (ISO 24622-1:2015). This design offers a common syntactic basis, but also helps maximize the semantic interoperability of CMDI-based metadata schemes. In the past, however, CMDI has not been used sufficiently to achieve its full potential toward semantic interoperability. With the advent of the Semantic Web and the idea of Linked Data, it is clear that CMDI's interoperability claim is currently limited to the CLARIN universe and that data sharing with other communities remains an issue that needs to be addressed.

In this chapter, we discuss steps toward extending CMDI's semantic interoperability beyond the Social Sciences and Humanities: We stress the need for an initial data curation step, in part supported by a relation registry that helps impose some structure on CMDI vocabulary; we describe the use of authority file information and other controlled vocabulary to help connecting CMDI-based metadata to existing Linked Data; we show how significant parts of CMDI-based metadata can be converted to bibliographic metadata standards and hence entered into library catalogs; and finally we describe first steps to convert CMDI-based metadata to RDF. The initial grassroots approach of CMDI (meaning that anybody can define metadata descriptors and components) mirrors the AAA slogan of the Semantic Web ("Anyone can say Anything about Any topic"). Ironically, this makes it hard to fully link CMDI-based metadata to other Semantic Web datasets. This paper discusses the challenges of this enterprise.

Motivation

CLARIN is a research infrastructure that enables Social Sciences and Humanities scholars to access and to process language-related resources and tools (Hinrichs and Krauwer 2014). CLARIN offers four types of services: (1) access to resources such as reference corpora, lexical resources, and grammars; (2) construction of virtual collections to combine resources that support the study of research questions; (3) deposition and archiving of resources to manage persistent access and citation; and (4) provision of web-based tools that help scholars in the analysis of textual data, such as taggers, named entity recognizers, geolocation tools, and the like. It is therefore essential to describe the research material consistently and conclusively to help users locate the data, evaluate its usefulness for the task at hand, and access the data. The description framework needs to be expressive to cater to the large variety of data types, interoperable to support the sharing of descriptions, sufficiently flexible to anticipate future technology changes, but also standardized enough to ensure that the framework is adopted and used by the communities. For these reasons, CLARIN has selected the Component MetaData Infrastructure (CMDI), which is an international standard (ISO 24622-1:2015). The first part of this chapter describes CMDI, highlights its design principles, and gives CMDI usage examples. The second part discusses the challenges to connect CMDI-based metadata to Linked Data.

A Distributed Infrastructure

The CLARIN infrastructure is a distributed network across various institutions and countries, rather than a centralized hub. This has both historical and practical reasons. Historically, individual institutions have grown their own ecosystems of repositories for resources and tools. The ecosystems' designs often differ in nature at organizational and technical levels so that they cannot be simply combined into a single, central, or overarching infrastructure. Besides the differences in the technical ecosystem of the institutions, resources at an institution often have strong license restrictions imposed on them so that such a resource (e.g., a newspaper corpus) cannot be accessed outside of the institution, or access to the resource can be subject to strict authentication and authorization procedures. Additionally, institutions tend to have their own research specializations, and hence very different types of research data and tools, along with different methodologies and technical requirements to access and work with them. It is thus better to maintain all research data under the auspices of the institution in charge that created the resource, rather than attempting to centralize the archiving at a central agency. Moreover, distributed infrastructures share the risk among the various stakeholders and institutions, distribute the task and cost for preservation, and therefore improve the sustainability of the infrastructure.

A Rich and Diverse Set of Language Resources

CLARIN offers a large variety of language resources, ranging from corpora with various (linguistic) annotations, lexical resources, psycholinguistic experiments, digital editions of books, spoken language recordings and their annotations, endangered languages documentation, and grammars to big data corpora that feed applications in the area of language technologies. The language resources come in many different languages, and while most resources are monolingual, many are bilingual or even multilingual, while others have many layers of annotation to support their study.

This variety poses a number of challenges that the research infrastructure must cope with. Given CLARIN's distributed architecture, most technical and organizational issues are addressed at the participating institutions. A unified cataloging of all resources, however, requires a central approach to harvest, understand, and harmonize their metadata descriptions. The metadata descriptions need to have a level of expressiveness that allows scholars to evaluate the relevance of the resources they describe. Here, descriptive categories such as resource title, creator, size, or language usually do not suffice to assess whether a given resource fits a scholar's needs. In fact, each of the resource types benefits from its own set of descriptive means. A lexical resource, for instance, needs to be described in terms of the number of lexical keywords/lexemes and definitions it contains, whereas such data categories are meaningless for, say, a text corpus. In the latter case, a scholar might care for a corpus's size (number of words), its language or languages, its type-token ratio, its genre, and so on. Also, the interpretation of the metadata depends on the context. A resource with a size of five megabytes is rather small when the resource is multimodal material, but rather large when it is a lexical resource. The provision of meaningful descriptions that help researchers to either safely disregard a resource or be prompted to investigate the resource further is an issue of utmost importance that any central access to a distributed infrastructure must address.

Describing Language Resources Adequately

In the library world, electronic resources are predominantly described with Dublin Core metadata (DC), a set of 15 descriptive categories, such as author, title, publisher, year, and copyright holder (ISO 15836-1:2017). In the area of language resources, the Open Language Archive Community proposed its own metadata set. It is based on the complete set of Dublin Core metadata terms,[1] yet the format allows the use of extensions to express community-specific qualifiers (Simons and Bird 2008). One type of language resource has its own encoding standard: Digital editions of text are often made available in terms of the Text

Encoding Initiative.[2] The Text Encoding Initiative (TEI) Guidelines for Electronic Text Encoding and Interchange include an extensive header that describes the resource with metadata. More-expressive metadata formats are available in the library world, such as MARC 21 (MARC-21 1999), and though all these descriptive schemas are both helpful and effective in their context, they lack expressiveness to describe the varying types of language-related resources.

There are three approaches to tackling this issue: (1) construct a rich set of metadata descriptors to form a single schema to describe all language-related research data; (2) construct multiple schemas, each catering to a type of language-related research data; and (3) construct smaller components to describe the various aspects of language-related resources and then combine them in a modular fashion to more complex schemas, one for each type of language-related resource. CLARIN follows the third approach.

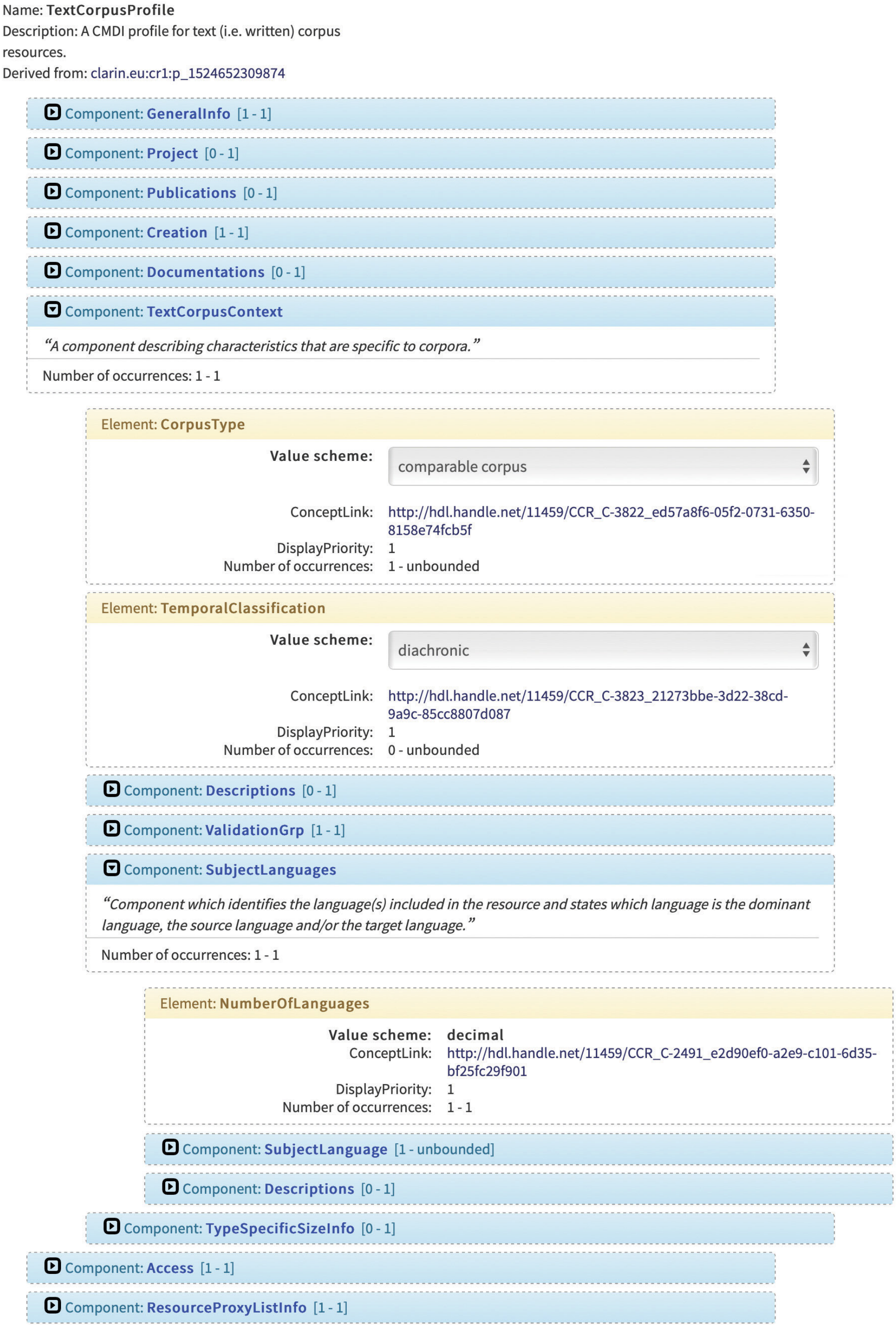
Component MetaData Infrastructure (CMDI)

Various types of language resources require different sets of metadata (i.e., profiles). In CMDI, a metadata profile for a given resource type is built by assembling prefabricated components, some of which are shared or reused across different schemas, while others are specific to the class of resource to be described. A CMDI component brackets elementary data descriptors or other, simpler components into a single unit. We obtain, thus, a hierarchical metadata system (Broeder et al. 2011).

Figure 6.1 shows a profile that can be used to describe text corpora. It has components /GeneralInfo/, /Project/, /Publications/, and /Creation/, among others, that capture information that is independent of the type of the language-related resource. The resource-specific component /TextCorpusContext/ makes use of the two data descriptors /CorpusType / and /TemporalClassification/. Each descriptor must have a value scheme specifying the type of its value and also must have a reference to its definition. In the given example, the ConceptLink has a handle reference, persistently addressing the corresponding element in the CLARIN concept registry (see below). Moreover, it is specified whether a data descriptor is optional or required, or whether it may occur multiple times.

The two data descriptors illustrate the expressive power required to adequately describe a resource of type corpus. The value scheme of /CorpusType/ can take a value from a predefined controlled vocabulary, which contains the terms “comparable corpus,” “parallel corpus,” “general corpus,” “reference corpus,” “learner corpus,” and so on. The controlled vocabulary for the element /TemporalClassification/ comprises the terms “diachronic,” “synchronic,” “historic,” “modern,” “other,” and “unknown.”

Each elementary data descriptor, or data category, should be defined in an external concept registry, such as the CLARIN concept registry (Schuurman, Windhouwer, Ohren, and Daniel 2016), or in its predecessor, the ISOcat registry, which is based on the ISO 12620:2009 standard, or should refer to other established metadata schemes, such as



Name: **TextCorpusProfile**
Description: A CMDI profile for text (i.e. written) corpus resources.
Derived from: clarin.eu:cr1:p_1524652309874

▶ Component: **GeneralInfo** [1 - 1]

▶ Component: **Project** [0 - 1]

▶ Component: **Publications** [0 - 1]

▶ Component: **Creation** [1 - 1]

▶ Component: **Documentations** [0 - 1]

▼ Component: **TextCorpusContext**

“A component describing characteristics that are specific to corpora.”

Number of occurrences: 1 - 1

Element: **CorpusType**

Value scheme: comparable corpus
ConceptLink: http://hdl.handle.net/11459/CCR_C-3822_ed57a8f6-05f2-0731-6350-8158e74fcb5f
DisplayPriority: 1
Number of occurrences: 1 - unbounded

Element: **TemporalClassification**

Value scheme: diachronic
ConceptLink: http://hdl.handle.net/11459/CCR_C-3823_21273bbe-3d22-38cd-9a9c-85cc8807d087
DisplayPriority: 1
Number of occurrences: 0 - unbounded

▶ Component: **Descriptions** [0 - 1]

▶ Component: **ValidationGrp** [1 - 1]

▼ Component: **SubjectLanguages**

“Component which identifies the language(s) included in the resource and states which language is the dominant language, the source language and/or the target language.”

Number of occurrences: 1 - 1

Element: **NumberOfLanguages**

Value scheme: decimal
ConceptLink: http://hdl.handle.net/11459/CCR_C-2491_e2d90ef0-a2e9-c101-6d35-bf25fc29f901
DisplayPriority: 1
Number of occurrences: 1 - 1

▶ Component: **SubjectLanguage** [1 - unbounded]

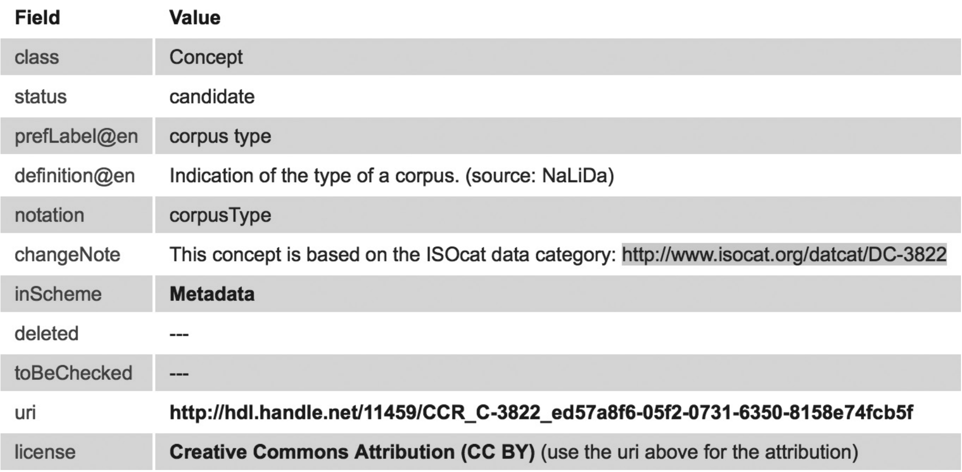
▶ Component: **Descriptions** [0 - 1]

▶ Component: **TypeSpecificSizeInfo** [0 - 1]

▶ Component: **Access** [1 - 1]

▶ Component: **ResourceProxyListInfo** [1 - 1]

Figure 6.1
The CMDI profile for a text corpus (screenshot from the Component Registry).



Field	Value
class	Concept
status	candidate
prefLabel@en	corpus type
definition@en	Indication of the type of a corpus. (source: NaLiDa)
notation	corpusType
changeNote	This concept is based on the ISOcat data category: http://www.isocat.org/datcat/DC-3822
inScheme	**Metadata**
deleted	---
toBeChecked	---
uri	**http://hdl.handle.net/11459/CCR_C-3822_ed57a8f6-05f2-0731-6350-8158e74fcb5f**
license	**Creative Commons Attribution (CC BY)** (use the uri above for the attribution)

Figure 6.2
The concept /corpus type/ (screenshot from the CLARIN Concept Registry).

the Dublin Core Metadata Set. Components and profiles are defined and stored centrally in the Component Registry (Ďurčo and Windhouwer 2014a). The components and profiles are defined using a Component Description Language; for each CMDI profile, the component registry can generate a corresponding XML schema definition (XSD).

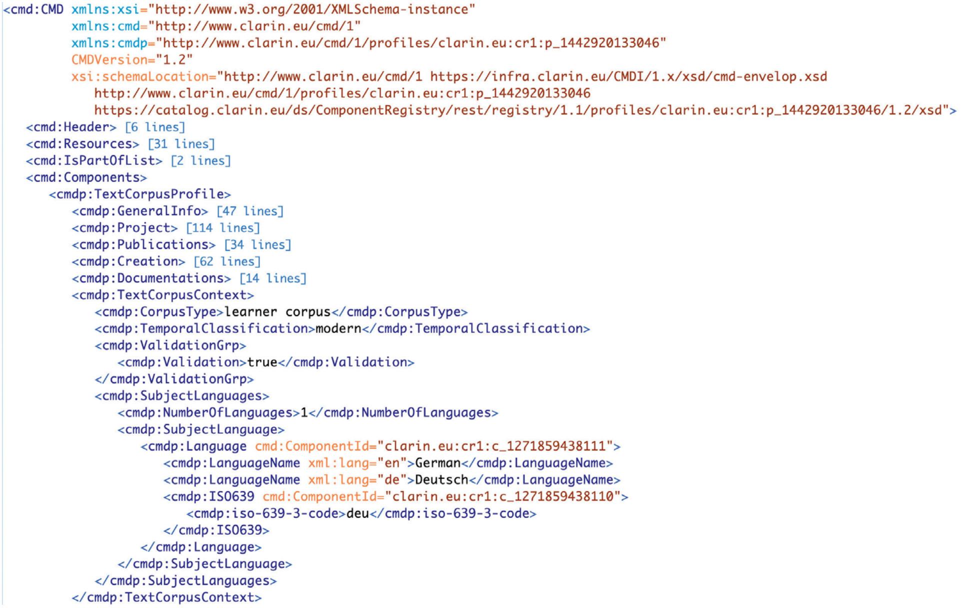
Figure 6.2 shows the concept /corpus type/ that is referenced from the profile for the description of text corpora. This concept is defined in the CLARIN concept registry, the most often used term registry in the CLARIN community.

In the CLARIN Component Registry, components can be searched for, edited, or newly created. The component registry has a public space that contains all components that have been published, which get a uniform and persistent ID so that others can use them, as well as a private space. In the latter, new components can be defined, and experimented with, before they may get published at a later stage.

Figure 6.3 shows the XML representation of a CMDI instance that describes a text corpus. Note that the instance refers to its profile in the xsi:schemaLocation attribute, and hence, standard XML technology can be used to validate whether the instance adheres to the schema.

The CLARIN infrastructure provides metadata modelers with a number of tools, such as:

- The CLARIN component registry (https://catalog.clarin.eu/ds/ComponentRegistry) and the CLARIN concept registry (https://openskos.meertens.knaw.nl/ccr/browser) for the definition and look-up of profiles, components, and data descriptors
- COMEDI (http://clarino.uib.no/comedi), a web-based editor for CMDI metadata
- SMC Browser, a web-based tool to visualize the hierarchical structure of CMDI profiles; see https://clarin.oeaw.ac.at/smc-browser/index.html



```
<cmd:CMD xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xmlns:cmd="http://www.clarin.eu/cmd/1"
         xmlns:cmdp="http://www.clarin.eu/cmd/1/profiles/clarin.eu:cr1:p_1442920133046"
         CMDVersion="1.2"
         xsi:schemaLocation="http://www.clarin.eu/cmd/1 https://infra.clarin.eu/CMDI/1.x/xsd/cmd-envelop.xsd
            http://www.clarin.eu/cmd/1/profiles/clarin.eu:cr1:p_1442920133046
            https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1442920133046/1.2/xsd">
   <cmd:Header> [6 lines]
   <cmd:Resources> [31 lines]
   <cmd:IsPartOfList> [2 lines]
   <cmd:Components>
      <cmdp:TextCorpusProfile>
         <cmdp:GeneralInfo> [47 lines]
         <cmdp:Project> [114 lines]
         <cmdp:Publications> [34 lines]
         <cmdp:Creation> [62 lines]
         <cmdp:Documentations> [14 lines]
         <cmdp:TextCorpusContext>
            <cmdp:CorpusType>learner corpus</cmdp:CorpusType>
            <cmdp:TemporalClassification>modern</cmdp:TemporalClassification>
            <cmdp:ValidationGrp>
               <cmdp:Validation>true</cmdp:Validation>
            </cmdp:ValidationGrp>
            <cmdp:SubjectLanguages>
               <cmdp:NumberOfLanguages>1</cmdp:NumberOfLanguages>
               <cmdp:SubjectLanguage>
                  <cmdp:Language cmd:ComponentId="clarin.eu:cr1:c_1271859438111">
                     <cmdp:LanguageName xml:lang="en">German</cmdp:LanguageName>
                     <cmdp:LanguageName xml:lang="de">Deutsch</cmdp:LanguageName>
                     <cmdp:ISO639 cmd:ComponentId="clarin.eu:cr1:c_1271859438110">
                        <cmdp:iso-639-3-code>deu</cmdp:iso-639-3-code>
                     </cmdp:ISO639>
                  </cmdp:Language>
               </cmdp:SubjectLanguage>
            </cmdp:SubjectLanguages>
         </cmdp:TextCorpusContext>
```

Figure 6.3
A CMDI instance for a language resource of type Text Corpus.

- CMDI2DC, a web service that converts CMDI-based profiles to Dublin Core (Zinn et al. 2016); see http://weblicht.sfs.uni-tuebingen.de/converter/Cmdi2DC/

The CLARIN center registry at https://centres.clarin.eu/oai_pmh maintains a list of all CLARIN repositories that provide their metadata publicly by using the Open Archive Initiative's Protocol for Metadata Harvesting (OAI-PMH). A central hub harvests all metadata at regular intervals and aggregates them into a single search index. The Virtual Language Observatory at http://vlo.clarin.eu/ enables users to explore the aggregated datasets via a dozen facets and to perform a full-text search on the metadata.

CMDI and Semantic Interoperability

In the past, the CLARIN community followed a grassroots approach to metadata management. The CMD infrastructure, in particular the registries, explicitly supported this movement. Anybody in the community was allowed and enabled to define metadata descriptors and components, mirroring the Semantic Web AAA slogan ("Anyone can say Anything about Any topic"). As a result, the registries were rapidly filled with descriptors to describe any possible aspect of a language resource. Often, users did not check whether an adequate descriptor or component already existed. Rather than using an existing one, new descriptors and components were defined helter-skelter. The effect of the grassroots movement is reflected by the content (many duplicates) and the size of the CLARIN registries. At the time of writing, the CLARIN Concept Registry provides over 3,000 entries; the CLARIN Component Registry has more than 1,000 public components and more than 180 public profiles.

It is clear that CMDI delivers on the grounds of syntactic interoperability. Based on XML, a CMDI instance documenting a resource is linked to a CMDI metadata schema, and XML validation is used to check whether the instance adheres to the schema. The main issue to address is the interpretation of the resulting syntactical structure. While most metadata elements are grounded in the CLARIN concept registry, the interpretation has to cope with the large number of duplicate entities being used and their varying contextual embedding.

The CLARIN Virtual Language Observatory (VLO) shows how to deal with this issue in an ad hoc manner. To deal with the large variety of different schemas, considerable parts of their data categories are semantically mapped to a dozen VLO search attributes.[3] Consider, for instance, the facet "language," which indexes all resources in terms of their language. In the CLARIN concept registry, there are at least four different entries that define the data descriptor "language" in some way or other. There are also CMDI components that refer to the Dublin Core element http://purl.org/dc/terms/language. All these entries are mapped to the facet "language," given that the data category is used in the "proper" context. If the data category is used in an "improper" context, say, to describe either the language of the resource's documentation or the native language of the resource's actor, the mapping will not take place.

While the mapping helps fix the issue for the VLO, the proliferation of duplicated data descriptors must be addressed by the CMDI community. In the future, CLARIN vocabulary must be far better managed. Users should use existing, established terms whenever possible, rather than defining their own set of terms. To alleviate the problem of contextual interpretation, when new terms need to be created, definitions should be specific rather than general. To minimize contextual interpretation, for instance, the descriptors /actorLanguage/ and /documentationLanguage/ should be preferred to simply /language/.

The CLARIN community has taken the first steps toward addressing the data curation issue. With the migration from the ISOcat registry to the SKOS-based concept registry, the grassroots approach to concept definition has been changed to a more controlled environment where designated members of the CLARIN community ("national CCR coordinators") now manage the vocabulary. Also, a best practices guide to metadata modeling within CMDI is currently being devised. On the software side, the SMC browser has been further developed to track the usage of profiles, components, and data descriptors across the CLARIN metadata set; it is a useful tool to support the curation of all existing content. Moreover, the CLARIN registries are being improved: The SKOS-based concept registry is easier to use than the ISO 12620:2009-based ISOcat registry, while the CLARIN component registry now adds a status to each component (one of Development, Production, and Deprecated).

Data curation in the CMDI universe, however, remains an enormous challenge. Any change in a CMDI profile (a change of a component or an elementary descriptor) must be mirrored by a corresponding update in all CMDI instances relying on the profile. With a million CMDI instances originating in 36+ CLARIN centers, this is a challenging task. Nevertheless, data curation must take place, and it shows that Semantic Web technology can support this process.

CMDI and Linked Data

The Semantic Web is built from structured data of uniform resource identifiers (URIs) that are highly interlinked. Berners-Lee (2006) defines Linked Data as accepting these four principles:

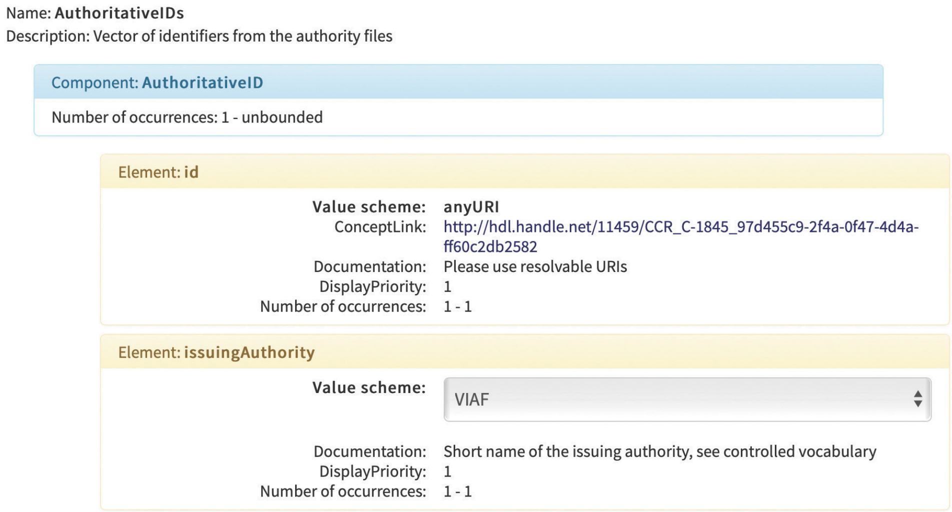
1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information using standards (Resource Description Framework [RDF[4]], SPARQL Protocol and RDF Query Language [SPARQL[5]]).
4. Include links to other URIs, so that they can discover more things.

With each CMDI profile and component in the CLARIN component registry, and each concept in the CLARIN concept registry being addressable with a persistent identifier, the CMD infrastructure clearly fulfills the first two conditions. The CLARIN community needs to take care of the remaining two conditions. For CMDI-based metadata to take part in Linked Data, it is necessary to add RDF support to the CMD infrastructure and to add links to existing datasets. For the latter, we need to map both the CMDI metadata vocabulary to existing Linked Data (LD) vocabulary and the value space of CMDI metadata to existing LD entities.

In CMDI-based metadata, there are a number of opportunities to link the value space of data descriptors with Linked Data entities. First and foremost, the names of resource creators, as well as the names of the institutions where the language resource originated or where it is hosted, should be linked to URIs that refer to names.

Authority File Information

Some CMDI metadata creators have started to use authority file information (Trippel and Zinn 2016). An authority file record gives the name of a person or institution a standardized representation. The authority file record links the standardized representation of a name with its alternative forms or spellings to accomplish the goal of disambiguation.



Name: **AuthoritativeIDs**
Description: Vector of identifiers from the authority files

Component: **AuthoritativeID**
Number of occurrences: 1 - unbounded

Element: **id**

Value scheme: **anyURI**
ConceptLink: http://hdl.handle.net/11459/CCR_C-1845_97d455c9-2f4a-0f47-4d4a-ff60c2db2582
Documentation: Please use resolvable URIs
DisplayPriority: 1
Number of occurrences: 1 - 1

Element: **issuingAuthority**

Value scheme: VIAF
Documentation: Short name of the issuing authority, see controlled vocabulary
DisplayPriority: 1
Number of occurrences: 1 - 1

Figure 6.4
A CMDI component for keeping authority file information.

Many libraries use authority files for identity management. The Integrated Authority File of the German National Library (GND) has about 11 million entries, which include over 7.5 million personal names and over 1 million names for corporate bodies (DNB 2016). The Virtual International Authority File at viaf.org is a joint project of more than 40 national libraries and is operated by the Online Computer Library Center. The aim of VIAF is to link together the national authority files of all project members to a single virtual authority file. Each VIAF record is associated with a URI and aggregates the information of the original authority records from the member states.

The International Standard Name Identifier (ISO 27729:2012) at http://isni.org/ holds nearly 9 million identities, including over 2.5 million names of researchers and more than 500,000 organization IDs. More recent initiatives include the Open Researcher and Contributor ID (ORCID) at orcid.org and ResearcherID at researcherid.com. All these authority agencies attach a uniform resource identifier to their records. Also note that many Wikipedia biographical articles refer to the URIs of the corresponding authority agencies. Their datasets are also prominent nodes in the Linked Open Data (LOD) project at http://linkeddata.org.

The flexibility of CMDI allows us to easily define a CMDI component to hold authority file information. Figure 6.4 shows the CMDI component /AuthoritativeIds/, which can be used to represent one or more authority records /AuthoritativeId/. Its first element, /id/, stores the persistent identifier of the authority agency, while /issuingAuthority/ refers to the agency that provides the data. All the aforementioned agencies can be selected as a value of this descriptor. The authors have amended all CMDI profiles that describe research data from the University of Tübingen to include this CMDI component for all data about persons and organizations. In practice, it shows that the majority of persons referred to in the CMDI metadata have an authority record. If not, persons can be asked to get an ORCID or ResearcherID. Also, most corporate bodies (such as university institutions and other research organizations) can be linked to such records. It shows that the GND from the German National Library is a good data source to link to, making it possible to uniquely identify, say, either the University of Tübingen or its linguistics department as the creator of many traditional publications (stemming from library catalogs), or to identify the research data (stemming from their repositories) it helped create. For more details, see Trippel and Zinn (2016). The authors hope that all CLARIN centers follow this example and add authority records to their data.

Use of Established Vocabularies

Any shared use of vocabulary supports Linked Data. The CLARIN Component Registry contains some components that make extensive use of externally defined metadata terms. The component /DcmiTerms/, for instance, gives a CMDI-based representation of all DCMI Metadata Terms.[6] To refer to languages, metadata providers can make use of the CMDI component /iso-language-639–3/ that represents the three-letter language codes as defined in the ISO 639-3:2007 code tables; see http://sil.org/iso639-3.[7] The CMDI component /Country/ makes available the country codes as defined by ISO 3166:2013.

The latter two components are outdated, because ISO has ended its support for URIs of the form https://cdb.iso.org/cdb to refer to language and country codes. While it is possible to use the URI http://sil.org/iso639-3/documentation.asp?id=deu (or http://www.lexvo.org/page/iso639-3/deu) to refer, say, to the ISO 639-3:2007 code for German (and to obtain more information about the referent), it is hard to say whether the links will remain resolvable 10 years from now. Therefore, the CLARIN community decided to import the ISO 639-3:2007 code set into CLAVAS, the newly created CLARIN vocabulary service based on OpenSKOS (http://openskos.org). This service aims to provide sustainable, persistent URIs to refer to ISO codes.[8]

Most CMDI data providers aim at using a controlled vocabulary to identify the media type of a resource, though referral to such data using the string datatype is often either incomplete or erroneous, and because users typically abstain from using persistent URIs. Explicit references to http://www.iana.org/assignments/media-types/media-types.xhtml are rare. Also, at the time of writing, CMDI metadata providers make little use of geographical databases such as geonames.org. Other missed opportunities to refer to shared vocabulary include the ISO 8601:2013 standard on dates and times, which is particularly interesting for the description of segments and their duration (intervals) from recordings, transcriptions, annotations, and the like. In the future, the CLAVAS vocabulary service may include the IANA terms and the other aforementioned terms to address this issue.

Link to Existing Vocabularies

There is ample potential to link CMDI-based data categories to the Semantic Web world, given that all descriptors in the CLARIN concept registry are addressable by persistent identifiers. To support data curation, semantic interoperability can be increased by relating CMDI-based data descriptors to each other. For this, reconsider the aforementioned mapping of CMDI data categories to facets to support faceted browsing in the Virtual Language Observatory. Here, the mapping is ad hoc rather than principled. In Windhouwer (2012) and Ďurčo and Windhouwer (2013), the authors propose establishing explicit ontological relationships between data descriptors. Using a new registry, the RELcat relation registry, it becomes possible to establish, for instance, owl:same-as or skos:exaxtMatch relations between semantically equivalent concepts, or to relate skos:closeMatch to almost semantically equivalent concepts. In the future, the CLARIN community must use RELcat to formalize the VLO mapping, and on the grand scale it must impose ontological insight onto the 3,000+ entries in the CLARIN concept registry.

Schema.org is an interesting ontology that CMDI metadata providers should consider using. Take the class http://schema.org/PostalAddress, for instance. It serves as an anchor point to address-related properties, most of which have near equivalents in the CLARIN concept registry: /locationAddress/, /locationRegion/, /locationCountry/, /locationContinent/, /email/, and /faxNumber/. The term /address/ could then be linked to PostalAddress, and the aforementioned terms to its properties. In Zinn, Hoppermann, and Trippel (2012), the authors propose mapping some of the entries of the CLARIN concept registry to the schema.org ontology, in part to increase CMDI's interoperability with regard to the Semantic Web community, and in part to support ongoing curation efforts within the CMDI community. So far, the CLARIN community has yet to discuss and agree on a mapping of CMDI vocabulary to schema.org or to vocabularies from other well-known concept registries or metadata schemes. Here, the RELcat relation strategy should be used to formalize the mapping described in Zinn, Hoppermann, and Trippel (2012) and to enter other term equivalencies. Community consensus could be marked by attaching a status to each mapping.

From CMDI to Linked Data via Bibliographic Metadata

Ongoing work is needed to move CMDI closer to the library world and subsequently toward the Semantic Web. In Zinn et al. (2016), the authors propose crosswalks (along with a web-based converter) between CMDI-based profiles and the library metadata standards Dublin Core and MARC 21. Having a CMDI-based record converted to MARC 21 helps its ingestion in the library catalog, but without authority information the new information is not linked to any prior information in the catalog (e.g., common author or common publisher), and hence is of limited use. With authority file information, we can link person-related metadata (in particular, the creator of a resource) with /dc:author/ information in bibliographic databases. This makes it possible to have a single entry point for the traditional publications of a researcher and for the research data he or she created. The same holds for institutions that help to create or host linguistic data and metadata.

The conversion of CMDI to Dublin Core comes with a significant information loss; the conversion from CMDI to MARC 21 preserves a considerable amount of information and is hence the preferred bibliographic format. Once a bibliographic format has been attained, there are existing converters to Semantic Web standards. From MARC 21, for instance, there is a mapping to RDF that can be used to generate RDF triples.[9]

From CMDI to RDF via Direct Conversion

In Ďurčo and Windhouwer (2014b), the authors propose a conversion from XML-based CMDI representations to RDF-based representations, addressing the third item in Berners-Lee's (2006) list. The conversion includes all levels of the CMD data domain: the CMD meta model as given in ISO 24622-1:2015, CMD profiles and component definitions, CMD concept definitions, and RDF representations for CMD instance data. In the future, the CLARIN component registry will offer RDF representations for all profiles and components. Here, the hierarchical representation of a CMD component will be represented by the component's URI (rooted in the CLARIN component registry) and by a dot-path to its subcomponents and elements. At the time of writing, the RDF conversion is ongoing development.

A true conversion of CMDI-based RDF data requires data sharing at the URI level; that is, CMDI-based metadata must make a healthy use of URIs to refer to persons, corporations, geographical places, and other web entities. The use of authority records in CMDI-based metadata descriptions strengthens the links to other datasets, but this can only be the first step. With the semantic mapping from CMDI vocabulary to existing Linked Data vocabulary still to be done—the RELcat registry needs to be filled with many more entries—it is hard to evaluate the adequacy of the CMDI to RDF conversion algorithm. Here, the community must gather more experience.

RDF is the lingua franca of Linked Open Data. The data format comes with RDF-based technology for storing or querying datasets. In terms of metadata management, however, RDF is less legible and harder to maintain. Clearly, the CLARIN infrastructure best supports the record-based CMDI, so this is the mandatory format today for all CLARIN data providers. To reap the benefits of Linked Open Data, all harvested data should be converted to RDF and made accessible through SPARQL endpoints. Given the distributed nature of the CLARIN infrastructure, such conversion will be done at the central hub once all harvested data are aggregated and harmonized. Here, the VLO is the best place for getting access to RDF representations of CMDI instances.

Discussion

The CLARIN community has taken initial steps toward achieving semantic interoperability with other communities and toward linking CMDI-based metadata with metadata available elsewhere. The large number of vocabularies available in the metadata world, however, seems to complicate the community's work. Clearly, the CMDI community must first face the challenge of curating its own datasets. Given the distributed nature of

CLARIN, a considerable part of this task must be tackled by the individual data providers. Each of them profits, however, from the CMD infrastructure, in terms of the following: an improved CLARIN Concept Registry, where national CCR coordinators are now in charge to manage (and to curate) all terms; a CLARIN component registry that will need to offer (and to better advertise) prefabricated components that are semantically grounded in the CCR and other term registries; and an evolving CLARIN RELcat relation registry, where CCR terms can be ontologically linked both to each other and to external vocabularies. With powerful tool support (the SMC browser), data curation should be taken seriously; manpower should be made available to address the curation and interoperability challenges in a timely manner.

The advent of the Semantic Web and the idea of Linked Data offer motivation as well as conceptual and technological support for this task. While data curation starts at home, it must not be limited to CLARIN's own backyard. There exists a plethora of vocabularies in the metadata universe, and it is often unclear which ones are best to use and how to use them effectively. Which of the vocabularies will persist through time, or at least through a number of decades? While ISO standards are good candidates, the authors have already experienced that URIs to them become unresolvable, which in turn provides a convincing argument to set up one's own terminology service (such as CLAVAS).

In Cole, Han, Weathers, and Joyner (2013), the authors also mention the challenge of "too many semantic options available for creating RDF representations" in the library context: early adopters of library LOD often have "developed their own namespaces and semantics when publishing their catalogue records as LOD data sets." As a result, the authors continue, "there are too many sets used for library LOD data sets. No single semantic net seems sufficient for describing library bibliographic data records." With the CMDI community moving closer to the Linked Data world, we may reach a similar conclusion with regard to metadata records on language-related resources and tools. Here, the CLARIN members should take into account the best practices that are currently being designed for transforming bibliographic metadata into Linked Data (see, for instance, Southwick 2015).

In this regard, it is worth emphasizing that the library world is also transitioning to the Semantic Web and Linked Data. The Library of Congress has proposed a new standard for library resource description called BIBFRAME, https://www.loc.gov/bibframe/. While the Library acknowledges the existence of different vocabularies (schema.org, Resource Description and Access [RDA], http://www.rda-rsc.org), the BIBFRAME vocabulary comes with its own namespace.[10] The authors of the bibliographic framework by the Library of Congress (2012) acknowledge that "the recommendation of a singular namespace is counter to several current Linked Data bibliographic efforts." However, they continue, "it is crucial to clarify responsibility and authority behind the schematic framework of BIBFRAME in order to minimize confusion and reduce the complexity of the resulting data formats. It will be the role of the Library's standards stakeholders to maintain the connections between BIBFRAME model elements and source vocabularies such as Dublin Core, FOAF, SKOS, and future, related vocabularies that may be developed to support different aspects of the Library workflow." The CLARIN community may well decide to follow the example of a singular namespace and to use the RELcat relation registry to link, whenever possible, to relevant vocabularies that play an important role in prominent Linked Data sets.

Conclusion

In the past, research data were hardly accessible. They resided on recording reel-to-reel tapes, floppy disks, or hard drives, and to gain access to data it was often necessary to contact the researcher who collected the data in the first place to learn details about the data (for free) or to make a copy of the requested material. Some institutions followed a systematic approach to collecting and archiving research data, and also devised their own metadata format to help describing and accessing the data. With many different archives devising their proprietary metadata language, it was hard, if not impossible, for researchers to search across collections. The CMDI framework for metadata aims at fostering syntactic and semantic interoperability. It enables archive maintainers and other users to first define their descriptive vocabulary in concept registries (or, whenever possible, to use the vocabulary defined there). With the basic vocabulary in place, larger metadata chunks, or "components," can be defined and made available to others in the CLARIN component registry. This grassroots movement helped archives to replace their proprietary description framework with CMDI-based descriptions. These research data are regularly harvested from the many different data providers at a central place. Following a curation and mapping phase, the data are entered into the Virtual Language Observatory, which in turn allows users to perform a faceted-based search to large aggregations of language-related material of an enormous variety.

By now, CMDI-based metadata have become the standard framework to describe language-related resources in the CLARIN community. The CMDI community still must address a number of issues. First, the CLARIN concept and component registries need better curation. In both registries, unused entries should be removed. To deal with duplicates and near duplicates, data descriptors should be interlinked with each other to state their ontological relationships. Also, their relationships with terms from other metadata schemes should be made explicit. Here, the RELcat relation registry, which was first introduced in Windhouwer (2012), should be finally released and effectively used for this purpose. Second, where applicable, vocabulary from established metadata schemes, such as Dublin Core, should be used to describe aspects about a resource that are independent from its type. With the usage of established vocabulary rather than CMDI homegrown terms, there is no need for a mapping. Third, for values of descriptors, authority files from the library world should be used whenever possible, as should ISO-based closed vocabularies for countries, languages, dates, and so on. Here, the CLAVAS vocabulary service should be used whenever possible, also to ensure the persistence of all URIs. Fourth, it is desirable that RDF become an integral part of the CLARIN infrastructure. Both the component and the concept registries should offer RDF exports and SPARQL endpoints for their entries. Also, the VLO resource viewer should make available RDF-based representations of the resources' metadata. With these adoptions, CMDI-based metadata can be easily linked to library catalogs and Linked Data, so that in the midterm as many as a million records describing language-related resources can finally become a highly interlinked part of the Linked Data cloud.

Notes

1. See http://dublincore.org/documents/dcmi-terms/.
2. See http://www.tei-c.org/.
3. See https://lux17.mpi.nl/isocat/clarin/vlo/mapping/index.html.
4. See https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.
5. See https://www.w3.org/TR/sparql11-query/.
6. See http://dublincore.org/documents/dcmi-terms/.
7. The CMDI component /iso-language-639–3/ is identified with the URI http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/components/clarin.eu:cr1:c_1271859438110.
8. The URI http://clavas.clarin.eu/clavas/public/api/concept/2bfb9f9a-e088-4473-bf8e-5de7b81e716f, for instance, refers to "deu," German.
9. See https://wiki.dnb.de/pages/viewpage.action?pageId=124132496.
10. The namespace for the BIBFRAME 2.0 vocabulary is http://id.loc.gov/ontologies/bibframe.

References

Berners-Lee, Tim. 2006. "Linked Data." Last modified June 18, 2009, accessed May 22, 2017. https://www.w3.org/DesignIssues/LinkedData.html.

Broeder, Daan, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. "A Pragmatic Approach to XML Interoperability—the Component Metadata Infrastructure (CMDI)." Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2–5, 2011.

Cole, Timothy W., Myung-Ja Han, William Fletcher Weathers, and Eric Joyner. 2013. "Library Marc Records into Linked Open Data: Challenges and Opportunities." *Journal of Library Metadata* 13 (2–3):163–196.

DNB. 2016. "Integrated Authority File (GND)." Deutsche Nationalbibliothek. Last modified October 21, 2016, accessed May 22, 2017. http://www.dnb.de/gnd.

Ďurčo, Matej, and Menzo Windhouwer. 2013. "Semantic Mapping in CLARIN Component Metadata. In: Metadata and Semantics Research." Seventh Metadata and Semantics Research Conference, Thessaloniki, Greece.

Ďurčo, Matej, and Menzo Windhouwer. 2014a. "The CMD Cloud." Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

Ďurčo, Matej, and Menzo Windhouwer. 2014b. “From CLARIN Component Metadata to Linked Open Data.” Ninth International Conference on Language Resources and Evaluation (LREC’14), Reykjavik, Iceland.

Hinrichs, Erhard, and Steven Krauwer. 2014. “The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars.” Ninth International Conference on Language Resources and Evaluation (LREC’14), Reykjavik, Iceland.

ISO 639-3:2007. Codes for the representation of names of languages—Part 3: Alpha-3 code for comprehensive coverage of languages. Geneva: International Organization for Standardization.

ISO 3166-1:2013. Codes for the representation of names of countries and their subdivisions—Part 1: Country codes. Geneva: International Organization for Standardization.

ISO 8601-1:2019. Date and time—Representations for information interchange—Part 1: Basic rules. Geneva: International Organization for Standardization.

ISO 12620:2009. Terminology and other language and content resources—Specification of data categories and management of a Data Category Registry for language resources. Withdrawn, Geneva: International Organization for Standardization.

ISO 15836-1:2017. Information and documentation—The Dublin Core metadata element set. Part 1: Core elements. Geneva: International Organization for Standardization.

ISO 24622-1:2015. Language resource management—Component Metadata Infrastructure (CMDI). Part 1: The Component Metadata Model. Geneva: International Organization for Standardization.

ISO 27729:2012. Information and documentation—International standard name identifier (ISNI). Geneva: International Organization for Standardization.

Library of Congress. 2012. “Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services.” http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.

MARC-21. 1999. “MARC 21 Format for Bibliographic Data.” Last modified update no. 24 (May 2017), accessed May 20, 2017. https://www.loc.gov/marc/bibliographic/.

Schuurman, Ineke, Menzo Windhouwer, Oddrun Ohren, and Zeman Daniel. 2016. “CLARIN Concept Registry: The New Semantic Registry.” *Linköping University Electronic Press* (123): 62–70.

Simons, Gary, and Steven Bird. 2008. “OLAC Metadata.” Last modified May 31, 2008, accessed May 20, 2017. http://www.language-archives.org/OLAC/metadata.html.

Southwick, Silvia B. 2015. “A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies.” *Journal of Library Metadata* 15 (1): 1–35.

Trippel, Thorsten, and Claus Zinn. 2016. “Enhancing the Quality of Metadata by Using Authority Control.” Fifth Workshop on Linked Data in Linguistics, Portorož, Slovenia.

Windhouwer, Menzo. 2012. “RELcat: A Relation Registry for ISOcat Data Categories.” Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, May 23–25, 2012.

Zinn, Claus, Christina Hoppermann, and Thorsten Trippel. 2012. “The ISOcat Registry Reloaded.” The Semantic Web: Research and Applications (ESWC 2012).

Zinn, Claus, Thorsten Trippel, Steve Kaminski, and Emanuel Dima. 2016. “Crosswalking from CMDI to Dublin Core and MARC 21.” Tenth International Conference on Language Resources and Evaluation (LREC’16), Portorož, Slovenia.

7 Expressing Language Resource Metadata as Linked Data: The Case of the Open Language Archives Community

Gary F. Simons and Steven Bird

Introduction

The Open Language Archives Community (OLAC) is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources.[1] The library is virtual because OLAC does not hold any of the resources itself, rather it aggregates a union catalog of all the resources held by the participating institutions. A major achievement of the community has been to develop standards for expressing and exchanging the metadata records that describe the holdings of an archive. Since its founding in 2000, the OLAC virtual library has grown to include over 300,000 language resources housed in 60 participating archives.[2] Because all the participating archives describe their resources using a common format and shared vocabularies, OLAC is able to promote discovery of these resources through faceted search across the collections of all 60 archives.[3]

The OLAC metadata standard prescribes an interchange format that uses a community-specific XML markup schema. In the meantime, Linked Data has emerged as a common data representation that allows information from disparate communities to be linked into an interoperating universal Web of Data. This chapter explores the application of Linked Data to the problem of describing language resources in the context of OLAC. The first section sets the baseline by describing the OLAC metadata standard. The next section discusses Linked Data and how the existing OLAC standards and infrastructure measure up against the rules of Linked Data. The third section then describes how we have implemented the conversion of OLAC metadata records into resources within the Linked Data framework. Finally, the fourth section considers the impact on the OLAC infrastructure, including both changes that have already been implemented in order to bring the resources of OLAC's participating archives into the Linguistic Linked Open Data (LLOD) cloud (Chiarcos et al. 2013), as well as the potential of embracing Linked Data as the basis for a revised OLAC metadata standard.

The OLAC Metadata Standard

OLAC has created an infrastructure for the discovery and sharing of language resources (Simons and Bird 2003, 2008d). The infrastructure is built on three foundational standards: *OLAC Process* (Simons and Bird 2006), which defines the governance and standards process; *OLAC Metadata* (Simons and Bird 2008a), which defines the XML format used for the exchange of metadata records; and *OLAC Repositories* (Simons and Bird 2008b), which defines the requirements for implementing a metadata repository that can be harvested by an aggregator using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[4]

The OLAC metadata scheme (Bird and Simons 2004) is based on Dublin Core, which is a standard originally developed within the library community to address the cataloging of web resources. At its core, Dublin Core has 15 basic elements for describing a resource: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, and Type. To support greater precision in resource descriptions, this basic set has been developed into an enriched set of metadata terms (DCMI 2012) that can be used to further qualify these elements. The qualifications are of two kinds—refinements that provide more specific meanings for the elements themselves and encoding schemes (including controlled vocabularies) that provide for standardized ways of representing the values of the elements.

The OLAC metadata format is defined by a community-specific XML schema that follows the published guidelines for representing qualified Dublin Core in XML (Powell and Johnston 2003). In addition to supporting the encoding schemes defined by the Dublin Core Metadata Initiative, those guidelines provide a mechanism for further extension via the incorporation of application-specific encoding schemes. The OLAC community has used its standards process to define five metadata extensions (Bird and Simons 2003, Simons and Bird 2008c) that use controlled vocabularies specific to language resources:

- Subject Language, for identifying with precision (using a code from the ISO 639 standard)[5,6] which language a resource is about
- Linguistic Type, for classifying the structure of a resource as primary text, lexicon, or language description
- Linguistic Field, for specifying a relevant subfield of linguistics
- Discourse Type, for indicating the linguistic genre of the material.
- Role, for documenting the parts played by specific individuals and institutions in creating a resource

The following is a sample metadata record in the XML format prescribed by the *OLAC Metadata* standard as it has been published by the Lyon-Albuquerque Phonological Systems Database, or LAPSyD. The described resource provides information on the phonological inventory, syllable structures, and prosodic patterns of the Cape Verde Creole language. The example below shows the complete metadata record as it is returned in a GetRecord request of the OAI-PMH:

```
<oai:record xmlns:oai=http://www.openarchives.org/OAI/2.0/
        xmlns:olac=http://www.language-archives.org/OLAC/1.1/
        xmlns:dc=“http://purl.org/dc/elements/1.1/”
        xmlns:dcterms=“http://purl.org/dc/terms/”
        xmlns:xsi=“http://www.w3.org/2001/XMLSchema-instance”>
          <oai:header>
            <oai:identifier>oai:www.lapsyd.ddl.ish-lyon.cnrs
            .fr:src692</oai:identifier>
            <oai:datestamp>2009-10-07</oai:datestamp>
          </oai:header>
          <oai:metadata>
            <olac:olac>
            <dc:title>LAPSyD Online page for Cape Verde Creole,
            Santiago dialect</dc:title>
            <dc:description>This resource contains information about
            phonological
            inventories, tones, stress and syllabic structures
            </dc:description>
            <dcterms:modified xsi:type=“dcterms:W3CDTF”>2012-05-17
            </dcterms:modified>
            <dc:identifier xsi:type=“dcterms:URI”>http://www.lapsyd.ddl
            .ish-
            lyon.cnrs.fr/lapsyd/index.php?data=view&amp;code=692
            </dc:identifier>
            <dc:publisher xsi:type=“dcterms:URI”>www.lapsyd.ddl.ish
            -lyon.cnrs.fr
            </dc:publisher>
            <dcterms:license xsi:type=“dcterms:URI”>http://
            creativecommons.org/licenses/by-nc-nd/3.0/
            </dcterms:license >
            <dc:type xsi:type=“dcterms:DCMIType”>Dataset</dc:type>
            <dc:format xsi:type=“dcterms:IMT”>text/html</dc:format>
            <dc:contributor xsi:type=“olac:role” olac:
            code=“author”>Maddieson,
            Ian</dc:contributor>
            <dc:subject xsi:type=“olac:linguistic-field”
            olac:code=“phonology”/>
            <dc:subject xsi:type=“olac:linguistic-field”
            olac:code=“typology”/>
            <dc:type xsi:type=“olac:linguistic-type”
            olac:code=“language _ description”/>
            <dc:language xsi:type=“olac:language” olac:code=“eng”/>
            <dc:subject xsi:type=“olac:language” olac:code=“kea”>Cape
            Verde Creole,
            Santiago dialect</dc:subject>
          </olac:olac>
</oai:metadata>
</oai:record>
```

In the example, we can see the basic features of OLAC metadata. Metadata elements come from the 15 elements of the basic dc namespace, plus the additional refinements from the dcterms namespace. The xsi:type attribute is used to declare the encoding scheme that is used to express a value precisely. When the encoded value comes from a controlled vocabulary that is enumerated in one of the OLAC recommendations listed above, the olac:code attribute is used to encode the value. In that case, the element content can optionally be used to express the denotation more specifically. For instance, the final element in the example above illustrates using an ISO 639-3 code to identify the language and adding a note to say more specifically that the resource pertains to a particular dialect.

Enter Linked Data

When OLAC began, developing purpose-specific XML markup for information interchange was a best current practice. In the intervening years, Linked Data (Berners-Lee 2006; Bizer, Heath, and Berners-Lee 2009) has emerged from the Semantic Web[7] activity of the World Wide Web Consortium as a strategy for linking disparate purpose-specific datasets into a single interoperating global Web of Data. The impetus for reframing OLAC metadata in terms of Linked Data has come from two directions. The first is the general trajectory of the Dublin Core Metadata Initiative and the wider library community. Librarians are recognizing that Linked Data represents an opportunity for libraries to integrate their information resources with the wider web (see, for instance, Byrne and Goddard 2010). Whereas Dublin Core was initially conceived as a simple record format, a new best practice has emerged in which an abstract model[8] is used in defining application profiles[9] that provide semantic interoperability with other applications within the Linked Data framework (Baker 2012). There is perhaps no stronger evidence for a major trend toward Linked Data in cataloging than the BIBFRAME[10] initiative at the Library of Congress, which is building on the Linked Data model to develop a replacement for the

MARC standard (Miller et al. 2012). Players in the OAI-PMH world are also working with Linked Data (Haslhofer and Schandl 2008, 2010; Davison et al. 2013).

The second impetus has come from the application of the Linked Data framework to the linking of linguistic data and metadata (Chiarcos, Nordhoff, and Hellmann 2012). With the emergence of a Linguistic Linked Open Data cloud (Chiarcos et al. 2013), OLAC as a major source of linguistic metadata has been notable by its absence. The work described herein has therefore sought to rectify this gap by bringing OLAC into the cloud of Linked Data.

What does it take to link into the Web of Data? The Linked Data paradigm is based on four simple rules (Berners-Lee 2006):

1. Use uniform resource identifiers (URIs) to name (identify) things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide them with useful information using RDF and other Semantic Web standards.
4. Include links to other URIs so that users can discover more things.

These rules serve as the backdrop for the discussion in the next sections, which describe how OLAC resources have been expressed as Linked Data and how those expressions have been incorporated into the OLAC infrastructure.

As the rules indicate, the Linked Data paradigm is built on two foundational standards. The first is the Resource Description Framework (RDF),[11] which is a model for the representation and interchange of data that is semantically interoperable. The second is Uniform Resource Identifiers (URI),[12] which provide a syntax for the creation of globally unique names for things in the world (including concepts). The RDF approach to semantic representation can be summarized as follows. Information is expressed as a set of statements. Each statement is a triple consisting of a subject, a predicate, and an object. The subject is a resource that is named by a URI. The predicate is a URI that names a property. In the case of representing Dublin Core in RDF, the metadata elements (like Title, Date, Creator) become properties. The object may be another resource named by a URI or it may be a literal value. A set of statements forms a directed graph, in which the resources and literals are nodes and the properties are directed arcs from subject to object. The fact that any collection of RDF graphs can be merged into a single, large graph forms the basis for the interoperation across data sources within the Linked Data approach.

Expressing OLAC Metadata as Linked Data

OLAC is a source for information about three kinds of resources: the controlled vocabularies it has developed for language resource description, descriptions of the archives that participate in OLAC, and descriptions of the language resources those archives hold. The next three subsections describe how each of these is expressed as Linked Data. A final subsection considers the issue of personal and organizational names, which is an area in which the current solution is not yet in line with the rules of Linked Data.

Controlled Vocabularies

The *OLAC Metadata Usage Guidelines*[13] specify many best practices in terms of controlled vocabularies that should be used in representing the values of the metadata elements. To comply with the rules of Linked Data, those values need to be represented as URIs. All the controlled vocabularies that are specified as encoding schemes in Dublin Core (such as DCMI Type and MIME Type) already have URIs and RDF descriptions in common use. This includes the ISO 639-1 and ISO 639-2 standards for language identification, which are implemented at the Linked Data Service[14] of the Library of Congress. For instance, the 639-2 code [deu] for German is represented by *http://id.loc.gov/vocabulary/iso639-2/deu*. Work is in progress to implement the entire ISO 639-3 code set in the same way at the LC Linked Data Service; in the meantime, we are using *lexvo.org* URIs—for example, *http://lexvo.org/id/iso639-3/deu*.

The four controlled vocabularies defined by OLAC itself (Linguistic Type, Linguistic Field, Discourse Type, and Role) were not previously implemented in RDF. These have now been implemented as hash namespaces, so that “lexicon” from the Linguistic Type vocabulary is now represented by *http://www.language-archives.org/vocabulary/type#lexicon*. The vocabularies are implemented in RDF by means of the Simple Knowledge Organization System (SKOS).[15] The vocabulary as a whole is first defined as an instance of a concept scheme. For instance, the following is the definition of the Linguistic Type scheme. This RDF sample (as are all the samples that follow) is expressed in the N3[16] syntax. The first line is a complete subject-predicate-object triple in which “a” is shorthand for the property rdf:type. A semicolon indicates that the next line will be another predicate-object pair for the same subject, whereas a comma indicates an additional object for the same subject and predicate:

```
<http://www.language-archives.org/vocabulary/type>
        a skos:ConceptScheme ;
        dc:title "OLAC Linguistic Data Type Vocabulary" ;
        dc:description "This document specifies the codes, or
        controlled vocabulary, for the Linguistic Data Type exten-
        sion of the DCMI Type element. These codes describe the
        content of a resource from the standpoint of recognized
        structural types of linguistic information." ;
        dc:publisher "Open Language Archives Community" ;
        dcterms:issued "2006-04-06" ;
```

```
          rdfs:isDefinedBy <http://www.language-archives.org/REC/type
          .html>, <http://www.language-archives.org/vocabulary/type
          .rdf> ;
          skos:hasTopConcept
            <http://www.language-archives.org/vocabulary/type
            #language _ description>,
            <http://www.language-archives.org/vocabulary/type
            #lexicon>,
            <http://www.language-archives.org/vocabulary/type
            #primary _ text> .
```

Each term in the vocabulary is then defined as a SKOS concept by mapping the definition, examples, and comments from the published vocabulary documentation[17] into the appropriate SKOS properties. Here is the definition of the term “lexicon”:

```
<http://www.language-archives.org/vocabulary/type#lexicon>
          a skos:Concept ;
          skos:inScheme <http://www.language-archives.org/vocabulary
          /type> ;
          skos:prefLabel "Lexicon" ;
          skos:definition "The resource includes a systematic listing
          of lexical items." ;
          skos:example "Examples include word lists (including com-
          parative word lists), thesauri, wordnets, framenets, and
          dictionaries, including specialized dictionaries such as
          bilingual and multilingual dictionaries, dictionaries of
          terminology, and dictionaries of proper names. Non-word-
          based examples include phrasal lexicons and lexicons of
          intonational tunes." ;
          skos:scopeNote "Lexicon may be used to describe any
          resource which includes a systematic listing of lexical
          items. Each lexical item may, but need not, be accompanied
          by a definition, a description of the referent (in the
          case of proper names), or an indication of the item's
          semantic relationship to other lexical items." .
```

In the case of the Linguistic Type, Linguistic Field, and Discourse Type vocabularies, the terms are concepts that serve as the values of metadata properties. In the case of the Role vocabulary, the terms of the vocabulary are properties themselves. More specifically, they are refinements of the dc:contributor property. The implementation of those terms adds that declaration.

Archive Descriptions

OLAC publishes an index of all participating archives[18] that links to a description of each archive. By virtue of building on the OAI-PMH, every archive has been assigned a unique identifier from the outset, and these are mapped to HTTP URIs to provide a location for the archive description. For instance, the HTTP URI for the LAPSyD archive that is the source of the sample OLAC metadata record given above is *http://www.language-archives.org/archive/www.lapsyd.ddl.ish-lyon.cnrs.fr.* Thus, with respect to archive descriptions, OLAC already complied with the first two rules of Linked Data. But as far as the third rule is concerned, an RDF form of the description was missing.

The OLAC archive description is a mandatory component of an OLAC metadata repository.[19] It was already assigned a namespace and was defined by an XML schema.[20] Providing an RDF rendering of the archive descriptions involved first creating an RDF schema[21] that defines the properties of an OLAC archive description and then implementing an XSLT script that transforms the archive description as harvested from the repository into the RDF equivalent. For example, the following is the RDF description of the LAPSyD archive

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix olac-archive: <http://www.language-archives.org/OLAC/1.1/olac
-archive#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
<http://www.language-archives.org/archive/www.lapsyd.ddl.ish-lyon.cnrs
.fr> a rdfs:Resource ;
         dc:title "Lyon-Albuquerque Phonological Systems Database
         (LAPSyD)" ;
         olac-archive:archiveURL <http://www.lapsyd.ddl.ish-lyon.cnrs
         .fr/lapsyd/> ;
         dc:contributor "Flavier, Sébastien (Developer)",
           "Maddieson, Ian (Creator)",
           "Marsico, Egidio (Editor)",
           "Pellegrino, François (Editor)" ;
         olac-archive:institution "CNRS and University of New Mex-
         ico" ;
         olac-archive:shortLocation "Lyon, FRANCE" ;
         olac-archive:synopsis "This OAI/OLAC metadata repository
         gives a metadata record for every language entry in the
         Lyon-Albuquerque Phonological Systems Database (LAPSyD)
         database. LAPSyD is a searchable database which provides
         phonological information (inventories, syllable structure
         and prosodic patterns) on a wide sample of the world's
         languages." ;
```

```
          olac-archive:access "Each language entry described in this
          repository is a public Web page that may be accessed with-
          out restriction. Reuse of material on the site is subject to
          the Terms of Use shown on the LAPSyD site." .
```

Language Resource Descriptions

Similarly for language resource descriptions, each language resource has always been identified by an HTTP URI, but an RDF form of the description was missing. Another XSLT script has been implemented to transform the OAI-PMH GetRecord response into an RDF equivalent. For instance, this process outputs the sample OLAC metadata record given above as the following RDF statements:

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix olac-field: <http://www.language-archives.org/vocabulary
/field#>.
@prefix olac-role: <http://www.language-archives.org/vocabulary/role
#>.
@prefix olac-type: <http://www.language-archives.org/vocabulary/type
#>.

<http://www.language-archives.org/item/oai:www.lapsyd.ddl.ish-lyon.cnrs
.fr:src692>
          a rdfs:Resource ;
          dc:publisher <http://www.language-archives.org/archive/www
          .lapsyd.ddl.ish-lyon.cnrs.fr> ;
          dc:title "LAPSyD Online page for Cape Verde Creole, San-
          tiago dialect" ;
          dc:description "This resource contains information about
          phonological inventories, tones, stress and syllabic struc-
          tures" ;
          dcterms:modified "2012-05-17"^^dcterms:W3CDTF ;
          dc:identifier <http://www.lapsyd.ddl.ish-lyon.cnrs.fr/lapsyd
          /index.php?data=view&code=692> ;
          dc:publisher <www.lapsyd.ddl.ish-lyon.cnrs.fr> ;
          dcterms:license <http://creativecommons.org/licenses/by-nc
          -nd/3.0/> ;
          dc:type <http://purl.org/dc/dcmitype/Dataset> ;
          dc:format <http://purl.org/NET/mediatypes/text/html> ;
          olac-role:author "Maddieson, Ian" ;
```

```
dc:subject olac-field:phonology, olac-field:typology ;
dc:type olac-type:language _ description ;
dc:language <http://lexvo.org/id/iso639-3/eng> ;
dc:subject <http://lexvo.org/id/iso639-3/kea>,
  "Note for [kea]: Cape Verde Creole, Santiago dialect" .
```

Note that in the first dc:publisher statement, the LAPSyD archive (as described in the RDF snippet in the preceding subsection) is declared to be the publisher of the metadata record. This is an application of the fourth rule of Linked Data in which the objects of the RDF statements should link to other URIs so that users can discover more things. The use of OLAC-specific vocabularies is seen beginning with the olac:author property, which comes from the OLAC Role vocabulary. In the next two statements, the property values come from the OLAC Field and OLAC Type vocabularies, respectively. A final feature of note is in the final statement that describes the subject language of the resource. In the OLAC metadata standard, first the subject language is identified by a code from ISO 639-3 as the value of the olac:code attribute and then free text may be added in the element content to give greater detail. This is translated into two RDF statements, one with an HTTP URI as the value and the other with a literal string as the value. In generating the latter, which is a comment for human consumption, the conversion process prepends "Note for [kea]:" to identify which ISO 639-3 language the comment is about.

The Problem of Personal Names

Having implemented the conversions described above, OLAC is now expressing language resource metadata as Linked Data. There is one respect, however, in which the results still fall short of the spirit of Linked Data in that they fail to comply with the fourth rule of Linked Data: "Include links to other URIs so that users can discover more things." The problem area is the use of literal strings to represent the names of persons who are contributors to the language resource. In the specific case of the resource description above, the user should be able to follow a URI to find out who "Maddieson, Ian" is.

For the practice of Linked Data across a general audience, the URI of a person's article in the English Wikipedia is a popular source of URIs for persons. Even better for Linked Data purposes is the corresponding URI from DBpedia,[22] which maps each Wikipedia article into an RDF resource. Within the library cataloging world, the gold standard is to use an identifier from a national library's authority file—as, for instance, the Library of Congress Name Authority File.[23] In this particular case, Ian Maddieson is a sufficiently eminent linguist that he can actually be found in both, though that will not be the case for the vast majority of people who contribute to language resources. An existing single source of URIs for over 34,000 persons across the field of linguistics is the Linguist List Directory of Linguists,[24] though these URIs are not ideal for use in Linked Data because they are not "Cool URIs."[25] Another source that provides even more URIs, but that lacks uniformity, is personal or professional home page URIs. The academic world has recognized the need to develop a standardized way of uniquely identifying those who have made contributions to the academic literature. In 2012 an open, nonprofit, community-based effort named ORCID (Open Researcher and Contributor ID)[26] was launched. In just four years, its registry has grown to include over 2.5 million unique researcher identifiers.

All the following are thus HTTP URIs that could be used to identify this particular author in a Linked Data context (though note that only dbpedia.org, id.loc.gov, and orcid.org comply with all four rules of Linked Data):

- https://en.wikipedia.org/wiki/Ian_Maddieson
- http://dbpedia.org/resource/Ian_Maddieson
- http://id.loc.gov/authorities/names/n84089547
- http://linguistlist.org/people/personal/get-personal-page2.cfm?PersonID=695
- http://www.unm.edu/~ianm/index.html
- http://linguistics.berkeley.edu/person/23
- http://orcid.org/0000-0002-0775-0555

At present the *OLAC Metadata Usage Guidelines*[27] recommend only that a contributor be identified "by means of a name in a form that is ready for sorting within an alphabetical index." Yet the OLAC infrastructure has no means of enforcing this guideline or even of ensuring that each contributor metadata element names only one contributor. As a result, in spite of providing a faceted search service[28] that offers interoperable search on 14 facets that have uniform metadata values across the community of archives, contributor is not one of those facets. This is an area in which the community will need to tighten its metadata guidelines and practices if it intends to support the identification of contributors both in Linked Data and in faceted search.

Incorporating Linked Data into the OLAC Infrastructure

OLAC has taken the first steps of incorporating Linked Data into its infrastructure. The new RDF vocabularies described earlier are in place, as are the RDF transformations for archive descriptions and language resource descriptions. The URIs for all these resources are configured following W3C best practices to support HTTP content negotiation[29] so that they return an HTML document by default, but return an RDF/XML document when the header of the HTTP request specifically asks for the application/rdf+xml MIME type. To contribute[30] to the cloud of Linguistic Linked Open Data (Chiarcos et al. 2013), the nightly metadata harvest creates a gzipped dump[31] of the RDF/XML rendering of every metadata record in the OLAC catalog, and that dataset has been registered at the DataHub[32] of the Open Knowledge Foundation.

Looking to the future, the OLAC metadata standard has not changed appreciably since version 1.0 was adopted in 2003. In light of the trend toward Linked Data in the wider metadata community, now may be a fitting time to develop a version 2.0 update that brings OLAC into line with Linked Data as well as other current best practices. Doing so would encourage the participating archives to create metadata that better interoperates with the global Web of Data. The open-endedness of the Linked Data approach would further allow archives to create even richer metadata by augmenting their resource descriptions with properties from any RDF vocabulary. Perhaps the greatest advantage would be the long-term benefit for the sustainability of the OLAC vision that could accrue from entering into the mainstream of library practices. However, there is a downside: Developing OLAC 2.0 would have a substantial cost in terms of requiring participating archives to reimplement their OLAC repositories.

One way forward would be to adopt a hybrid approach. The OLAC harvester could support both OLAC 1.1 and 2.0. All 2.0 metadata would be back translated into 1.1 format so that all existing services continue to work. By the same token, all 1.1 metadata would be forward translated into 2.0 format and fed into an RDF aggregator that could capture all the added richness of 2.0 metadata. OLAC could then begin to develop new services that take full advantage of the Linked Data paradigm, including offering semantic search over the OLAC catalog by providing an endpoint for SPARQL (the query language for RDF).[33]

Conclusion

Given the core values of the OLAC process, one of which is that decisions be made by consensus and that the greatest voice is given to those who are implementing the standards, updating the OLAC metadata standard to a new version based on Linked Data is not a step that can be taken lightly. Moving to OLAC 2.0 would be a major effort requiring the participating archives around the world both to agree and to reimplement. Still, the time is surely ripe for OLAC to consider such an update to its standards and infrastructure, particularly in light of the potential for a future in which its language resource descriptions could interoperate seamlessly with the wider library cataloging community—and even more broadly with the global Web of Data.

Notes

1. http://www.language-archives.org/.
2. http://www.language-archives.org/archives.
3. http://search.language-archives.org.
4. https://www.openarchives.org/pmh/.
5. http://www.loc.gov/standards/iso639-2/.
6. http://www.sil.org/iso639-3/.

7. http://www.w3.org/standards/semanticweb/.
8. http://dublincore.org/documents/abstract-model/.
9. http://dublincore.org/documents/profile-guidelines/.
10. http://www.loc.gov/bibframe/.
11. http://www.w3.org/RDF/.
12. http://www.w3.org/Addressing/.
13. http://www.language-archives.org/NOTE/usage.html.
14. http://id.loc.gov/.
15. http://www.w3.org/2004/02/skos/.
16. http://www.w3.org/TeamSubmission/n3/.
17. http://www.language-archives.org/REC/type.html.
18. http://www.language-archives.org/archives.
19. http://www.language-archives.org/OLAC/repositories.html#OLAC%20archive%20description.
20. http://www.language-archives.org/OLAC/1.1/olac-archive.xsd.
21. http://www.language-archives.org/OLAC/1.1/olac-archive.rdf.
22. http://wiki.dbpedia.org/.
23. http://id.loc.gov/authorities/names.html.
24. http://linguistlist.org/people/personal/.
25. http://www.w3.org/TR/cooluris/.
26. http://orcid.org/.
27. http://www.language-archives.org/NOTE/usage.html.
28. http://search.language-archives.org/.
29. http://www.w3.org/TR/swbp-vocab-pub/.
30. http://wiki.okfn.org/Working_Groups/Linguistics/How_to_contribute.
31. http://www.language-archives.org/static/olac-datahub.rdf.gz.
32. https://datahub.io/dataset/olac.
33. https://www.w3.org/TR/sparql11-overview/.

References

Baker, Thomas. 2012. "Libraries, Languages of Description, and Linked Data: A Dublin Core Perspective." *Library Hi Tech* 30 (1): 116–133.

Berners-Lee, Timothy. 2006. "Design Issues: Linked Data: World Wide Web Consortium." http://www.w3.org/DesignIssues/LinkedData.html.

Bird, Steven, and Gary F. Simons. 2003. "Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources." *Computers and the Humanities* 37 (4): 375–388.

Bird, Steven, and Gary F. Simons. 2004. "Building an Open Language Archives Community on the DC Foundation." In *Metadata in Practice*, edited by Diane I. Hillmann and Elaine L. Westbrooks, 203–222. Chicago: American Library Association.

Bizer, Christian, Thomas Heath, and Timothy Berners-Lee. 2009. "Linked Data—The Story So Far." *International Journal on Semantic Web and Information Systems* 5 (3): 1–22, doi:10.4018/jswis.2009081901.

Byrne, Gillian, and Lisa Goddard. 2010. "The Strongest Link: Libraries and Linked Data." *D-Lib Magazine* 16 (11): 5.

Chiarcos, Christian, Sebastian Nordhoff, and Sebastian Hellmann, eds. 2012. *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*. Heidelberg: Springer.

Chiarcos, Christian, Steven Moran, Pablo N. Mendes, Sebastian Nordhoff, and Robert Littauer. 2013. "Building a Linked Open Data Cloud of Linguistic Resources: Motivations and Developments." In *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, edited by I. Gurevych and J. Kim, 315–348. Berlin: Springer. doi:10.1007/978-3-642-35085-6_12.

Davison, Stephen, Yukari Sugiyama, Elizabeth McAulay, and Claudia Horning. 2013. "Enhancing an OAI-PMH Service Using Linked Data: A Report from the Sheet Music Consortium." *Journal of Library Metadata* 13 (2–3): 141–162. doi:10.1080/19386389.2013.826067.

DCMI. 2012. "DCMI Metadata Terms." Dublin Core Metadata Initiative. http://dublincore.org/documents/dcmi-terms/.

Haslhofer, Bernhard, and Bernhard Schandl. 2008. "The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data." In *Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008)*. https://eprints.cs.univie.ac.at/284/

Haslhofer, Bernhard, and Bernhard Schandl. 2010. "Interweaving OAI-PMH Data Sources with the Linked Data Cloud." *International Journal of Metadata, Semantics and Ontologies* 5 (1): 17–1. doi:10.1504/IJMSO.2010.032648.

Miller, Eric, Uche Ogbuji, Victoria Mueller, and Kathy MacDougall. 2012. *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services*. Washington, D.C.: Library of Congress. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf.

Powell, Andrew, and Peter Johnston. 2003. "Guidelines for Implementing Dublin Core in XML." Dublin Core Metadata Initiative. http://dublincore.org/documents/dc-xml-guidelines/.

Simons, Gary F., and Steven Bird. 2003. "The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources." *Literary and Linguistic Computing* 18 (2): 117–128.

Simons, Gary F., and Steven Bird. 2006. *OLAC Process*. Open Language Archives Community. http://www.language-archives.org/OLAC/processa.html.

Simons, Gary F., and Steven Bird. 2008a. *OLAC Metadata*. Open Language Archives Community. http://www.language-archives.org/OLAC/metadata.html.

Simons, Gary F., and Steven Bird. 2008b. *OLAC Repositories*. Open Language Archives Community. http://www.language-archives.org/OLAC/repositories.html.

Simons, Gary F., and Steven Bird. 2008c. *Recommended Metadata Extensions*. Open Language Archives Community. http://www.language-archives.org/REC/olac-extensions.html.

Simons, Gary F., and Steven Bird. 2008d. "Toward a Global Infrastructure for the Sustainability of Language Resources." In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, 20–22 November 2008, Cebu City, Philippines*, edited by R. Roxas, 87–100. Manila: De La Salle University.

8 TalkBank Resources for Psycholinguistic Analysis and Clinical Practice

Nan Bernstein Ratner and Brian MacWhinney

Introduction

The formation of a network of Linguistic Linked Open Data (LLOD) can contribute in many important ways to the advancement of the study of language structure, usage, processing, and acquisition. The chapters in this book present a comprehensive overview of various efforts to build this new structure. The current chapter will show how the TalkBank system has already succeeded in realizing many of these goals and could eventually support still others. TalkBank had its origins in 1985 with the Child Language Data Exchange System (CHILDES), founded by Brian MacWhinney and Catherine Snow; both the first and second author of this chapter continue to work to enlarge and maintain its growing resources.

TalkBank (https://talkbank.org) is now the largest open repository of data on spoken language. Initially, these data were represented primarily in transcript form. However, new TalkBank corpora now include linkages of transcripts to media (audio and video) on the utterance level, as well as extensive annotations for morphology, syntax, phonology, gesture, and other features of spoken language.

An important principle underlying the TalkBank approach is that all its data are transcribed in a single consistent format. This is the CHAT format (talkbank.org/manuals/chat.pdf), which is compatible with the CLAN programs (talkbank.org/manuals/clan.pdf). This format has been developed over the years to accommodate the needs of a wide range of research communities and disciplinary perspectives. Using conversion programs available inside CLAN, the CHAT format can be automatically converted both to and from the formats required for Praat (praat.org), Phon (phonbank.talkbank.org), ELAN (tla.mpi.nl/tools/elan), CoNLL (universaldependencies.org/format.html), ANVIL (anvil-software.org), EXMARaLDA (exmaralda.org), LIPP (ihsys.com), SALT (saltsoftware.com), LENA (lenafoundation.org), Transcriber (trans.sourceforge.net), and ANNIS (corpus-tools.org/ANNIS). For each of these conversions, the CHAT format recognizes a superset of information types (dates, speaker roles, intonational patterns, retrace markings, and so on). This means that, when data are converted into the other formats, there must always be a method for protecting data types not recognized in those programs against loss. This is done in two ways. First, users can often hide CHAT data in special comment fields that are not processed by the program but that will be available for export. Second, when employing the other programs, users must be careful not to alter codes in CHAT format that mark aspects that cannot be recognized by the other programs. There are no cases in which information created in the other programs cannot be represented in CHAT, because CHAT is a superset of the information represented in these other programs.

TalkBank is composed of a series of specialized language banks, all using the same transcription format and standards. These include CHILDES (https://childes.talkbank.org) for child language acquisition, AphasiaBank (https://aphasia.talkbank.org) for aphasia, PhonBank (https://phonbank.talkbank.org) for the study of phonological development, TBIBank (https://tbi.talkbank.org) for language in traumatic brain injury, DementiaBank (https://dementia.talkbank.org) for language in dementia, FluencyBank (https://fluency.talkbank.org) for the study of fluency development/disorder, HomeBank (https://homebank.talkbank.org) for daylong recordings in the home, CABank (https://ca.talkbank.org) for Conversation Analysis, SLABank (https://slabank.talkbank.org) for second language acquisition, ClassBank (https://class.talkbank.org) for studies of language in the classroom, BilingBank (https://biling.talkbank.org_ for the study of bilingualism and code-switching, LangBank for the study and learning of classical languages, SamtaleBank (https://samtalebank.talkbank.org) for Danish conversations, the SCOTUS corpus in CABank with 50 years of oral arguments linked to transcripts at the Supreme Court of the United States, and the spoken portion of the British National Corpus, also in CABank. We and our collaborators are continually adding corpora to each of these collections. The current size of the text database is 1.4TB and there are an additional 5TB of media data. All the data in TalkBank are freely open to downloading and analysis, with the exception of the data in AphasiaBank, HomeBank, and research data in FluencyBank, which are password protected. The CLAN program and the related morphosyntactic taggers are all free and open-sourced through GitHub (http://github.com).

These databases and programs have been used widely in the research literature. CHILDES, the oldest and most widely recognized of these databases, has been used in over 6,500 published articles. PhonBank has been used in 480 articles and AphasiaBank has been used in 212 publications. In general, the longer a database has been available to researchers, the more that its use has become integrated into the basic research methodology and publication history of the field.

Metadata for the transcripts and media in these various TalkBank databases have been entered into the two major systems for accessing linguistic data: OLAC (see Simons and Bird in this volume) and CMDI/TLA (see Trippel and Zinn, also in this volume). Each transcript and media file has been assigned a PID (permanent ID) using the Handle System (www.handle.net). In addition, each corpus has received a DOI (digital object identifier) code. The metadata available through these systems, along with the data in the individual files, implements each of the requirements of the DTA tool system (Blume et al., this volume). The PID numbers are encoded in the header lines of each transcript file and the DOI numbers are entered into HTML web pages that include extensive documentation for each corpus, photos and contact information for the contributors, and articles to be cited when using the data. All these resources are periodically synchronized using a set of programs that rely on the fact that there is a completely isomorphic hierarchical structure for the CHAT data, the XML versions of the CHAT data, the HTML web pages, and also the media files. If information is missing for any item within this parallel set of structures, the updating program reports the error and it is fixed. All this information is then published using an OAI-PMH (www.openarchives.org/pmh) compatible method for harvesting through systems such as the Virtual Language Observatory at https://vlo.clarin.eu (VLO) developed through the CLARIN initiative (https://clarin.eu).

For 10 of the languages in the database, we provide automatic morphosyntactic analysis using the MOR, POST, and MEGRASP programs built into CLAN. These languages are Cantonese, Chinese, Dutch, English, French, German, Hebrew, Japanese, Italian, and Spanish. Tagging is done by MOR, disambiguation by POST, and dependency analysis by MEGRASP. Details regarding the operation of the taggers, disambiguators, and dependency analyzers for these languages can be found in MacWhinney (2008). Processing in each of these languages involves differing computational challenges. The complexity and linguistic detail required for analysis of Hebrew forms is perhaps the most extensive. In German, special methods are used for achieving tight analysis of the elements of the noun phrase. In French, it is important to mark various patterns of suppletion in the verb. Japanese requires quite different codes for parts of speech and dependency relations. Eventually, the codes produced by these programs will be harmonized with the GOLD ontology (Langendoen in this volume). In addition, we compute a dependency grammar analysis for each of these 10 languages, which we will harmonize with the Universal Dependency tagset (https://universaldependencies.org).

Because these morphosyntactic analyzers all use a parallel technology and output format, CLAN commands can be applied to each of these 10 languages for uniform computation of indices such as MLU (mean length of utterance), vocd (vocabulary diversity), pause duration, and various measures of disfluency. In addition, we have automated language-specific measures such as DSS or Developmental Sentence Score (for English and Japanese) and IPSyn. Following the method of Lubetich and Sagae (2014), we are now developing language-general measures based on classifier analysis that can be applied to all 10 languages using the codes in the morphological and grammatical dependency analyses. However, there are many other languages in the database for which we do not yet have morphosyntactic taggers. This means that it is a priority to construct MOR systems for languages with large amounts of CHILDES and TalkBank data, such as Catalan, Dutch, Indonesian, Polish, Portuguese, and Thai.

Using these data and methods, researchers have been able to evaluate the use of different approaches to comparable data. Such comparisons have been particularly fruitful in studies of the acquisition of morphology and syntax. For example, the debate between connectionist models of learning and dual-route models focused on data regarding the learning of the English past tense (Marcus et al. 1992; Pinker and Prince 1988; MacWhinney and Leinbach

1991) and later on data from German plural formation (Clahsen and Rothweiler 1992). In syntax, emergentists (Pine and Lieven 1997) have used CHILDES data to elaborate an item-based theory of learning of the determiner category, whereas generativists (Valian, Solt, and Stewart 2009) have used the same data to argue for innate categories. Similarly, CHILDES data in support of the Optional Infinitive Hypothesis (Wexler 1998) have been analyzed in contrasting ways using the MOSAIC system (Freudenthal, Pine, and Gobet 2010) to demonstrate constraint-based inductive learning. In these debates, and many others, the availability of a shared open database has been crucial in the development of analysis and theory.

Through these various methods of transcript format conversion, metadata publication, grammatical analysis, and data sharing, TalkBank has already fulfilled many of the goals of the LLOD project. As a result of these efforts, TalkBank has been recognized as a Center in the CLARIN network (clarin.eu) and has received the Core Trust Seal (https://coretrustseal.org). TalkBank data have also been included in the SketchEngine corpus tool (http://sketchengine.co.uk).

However, there are other goals of the LLOD project that seem to be currently out of the reach of spoken language corpora like TalkBank. The type of linkage proposed by Chiarcos and colleagues (this volume) and perhaps even the LAPPS system (Ide, this volume) would require a major effort to cross-index the individual lexical or morphological items in the many TalkBank databases. Such linkage makes sense for lexical databases or coding systems, because these involve linkages that can directly yield secondary analyses. For example, linkages between WordNet systems (http://wordnet.princeton.edu) in various languages or grammatical coding features (Langendoen, this volume) can directly facilitate a variety of NLP (natural language processing) tasks, such as translation, tagging, metaphor analysis, and information extraction. However, the value of linkages between entities for spoken language corpora has yet to be demonstrated. For these corpora, the role of individual lexical items depends entirely on the overall syntactic and discourse context, and it is not clear how these relations can be evaluated through simple links on the lexical or featural level. For these resources, the most important analytic tools involve corpus-based searches, such as those available in the TalkBankDB system at https://talkbank.org/DB.

An additional problem facing the task of linkages across spoken language data arises from the fact that many data centers do not make their data publicly available. For example, the majority of the materials in The Language Archive (tla.mpi.nl) cannot be directly accessed, and many are not available for access at all. The materials collected by the Linguistic Data Consortium (ldc.org) are only available to subscribers, thereby making them off limits for linked open access. Of the major databases for spoken language data, only TalkBank provides completely open access to records in a consistent XML format. Thus, TalkBank would seem to be a good target for integration into the LLOD project, once methods for dealing with spoken language corpora have been developed.

Rather than focusing on LLOD linkages across spoken language corpora, TalkBank has developed other methods for between-corpus linkage. Two of these methods have already been discussed. The first method involves the construction of programs that can convert between CHAT format and formats used by other analytic programs. That work has largely been completed. The second method is the construction and publication of metadata to the VLO system for indexing corpora, transcripts, and media. This work, too, has mostly been completed.

We are now actively engaged in the development of a third approach to between-corpus linkage. This method permits automatic quantitative comparisons between corpora or subsections of a given corpus. The goal here is to be able to compare data from speakers at different ages, speaking different languages, in different tasks and situations, at different stages of learning, and with different clinical profiles. In the balance of this chapter, we will outline the development of one of these methods, called KIDEVAL, for comparing child language data. A parallel system, called EVAL, has also been developed for making comparisons across samples of speech from persons with aphasia (PWAs). The EVAL system makes use of the fact that the data in AphasiaBank were all collected with a single consistent protocol. Based on these protocol data, we can extract group means for individual aphasia types (Broca's, Wernicke's, anomia, global, transcortical motor, and transcortical sensory), which we can then use as comparisons for the results from individual PWAs. For child language data, we have identified a subset of the database that can be used in a similar way to make comparisons within age groups. Comparisons of this type are fundamental to the process of clinical assessment, as well as to the study of basic developmental processes.

Child Language Sample Analysis

For the assessment of child language abilities, language sample analysis (LSA) provides a very high degree of ecological validity and "authenticity," as mandated by current educational policies (Overton and Wren 2014). It supplements standardized assessment by providing a snapshot, as it were, of a given child's language "in action." More critically, it provides baseline insights into the child's strengths and weaknesses across the range of language skills necessary for age-appropriate communication, from vocabulary to syntax to pragmatics. These skills can be tracked in natural contexts over time (Price, Hendricks, and Cook 2010). LSA provides clinicians with tangible goals for therapy unlikely to emerge from results of standardized testing but that can be prioritized for intervention (Overton and Wren 2014). In the absence of norm-referenced assessments for children speaking non-mainstream dialects or English as a Second Language, LSA also can provide less biased and more informative information about a child's expressive language skills and needs (Caesar and Kohler 2007; Gorman 2010).

However, there are a number of practical issues in using LSA for clinical purposes that tend to diminish the frequency (and depth) of its use in actual clinical practice (Gorman 2010). While the self-reported use of LSA has been steadily climbing in reports from 1993 to 2000 (Hux 1993; Eisenberg, Fersko, and Lundgren 2001; Kemp and Klee 1997), most SLPs (Speech-Language Pathologists) report compiling relatively short samples in real-time notation and using them primarily to compute mean length of utterance (MLU; Price, Hendricks, and Cook 2010; Finestack and Satterlund 2018), despite the fact that MLU is not a good standalone measure for identifying language impairment (Eisenberg, Fersko, and Lundgren 2001). In addition, Lee and Canter (1971) found that less than one-third of respondents computed an additional measure, the most popular being DSS. Very recently, Finestack and Satterlund (2018) found that only about 30% of American SLPs compute "informal" language sample measures. Of these, from 86 to 94% (depending upon age of child) used MLU. Type-token ratio (TTR) was used by only about 25–32% of respondents. Use of DSS had fallen to roughly 15% of SLPs, and other measures were used by fewer than 10% of SLPs who conducted LSA.

It is well acknowledged that good LSA can be quite time-consuming (Overton and Wren 2014). Some studies have estimated that it takes up to 8 hours of training and from 45 minutes to one hour of work after a transcript has been generated to compute DSS (Long and Channell 2001; Cochran and Masterson 1995). One study (Gorman 2010) estimated that it takes more than 30 minutes per sample following transcription to compute the Index of Productive Syntax (IPSYN; Scarborough 1990). Hand computation of most LSA measures, even the time-honored MLU, is quite prone to error. It is difficult to use the same worksheet to compute multiple linguistic measures, and it is a waste of time to transfer handwritten scribbles of what the child said to most scoring protocols. Thus, even by self-report, LSA is not used by many clinicians, and is not intensively exploited by most to inform child language assessment. Those who do LSA often use a sample that is much too short to meet the intended sample size for the measures that are computed (Westerveld and Claessen 2014), sometimes 50–75% fewer utterances than recommended.

Computer-assisted LSA can solve all the problems listed above (time, accuracy and depth of analysis; Heilmann 2010; Price, Hendricks, and Cook 2010; Evans and Miller 1999; Miller 2001; Hassanali 2014), but is not very frequently used in practice. A recent study estimated that only 12.5% of SLPs in Australia use computer-assisted transcription and analysis (Westerveld and Claessen 2014), and there is little to suggest that their American counterparts use such procedures at a significantly higher rate (Price, Hendricks, and Cook 2010). Finestack and Satterlund (2018) recently found that computer-assisted LSA was used by only 1–5% of American SLPs. As we will suggest, use of computers to aid in sample transcription and analysis, particularly using free utilities such as CLAN that additionally link the sample to an audio- or video-recorded record of the child's actual speech sample, can greatly improve the speed, accuracy, and informativeness of language sample analysis and, by extension, can also aid in clinical assessment, therapy planning, and measurement of therapeutic progress.

In this chapter, we will illustrate the utility of LSA conducted using CLAN and the KIDEVAL utility that uses two separate datasets. The first is a large cohort of very young children followed as part of a single research study. The second is a review of data obtained from the CHILDES Project Archive that we use to evaluate the potential utility of certain LSA measures at particular ages. Many LSA measures lack robust normative or comparison reference values, therefore the data in CHILDES can greatly augment what we currently know through measures such as MLU, DSS, IPSYN, VOCD, and others.

KIDEVAL in Action

In this section, we summarize how we have used the KIDEVAL utility to assess the dyadic interactions of a large cohort of infants and their mothers ($n = 125$), who were sampled at 7, 10, 11, 18, and 24 months as part of a larger study examining possible predictors of later child language skills (Newman, Rowe, and Ratner 2015). The scope of the project was quite daunting: We had ~125 families and conducted 5 play sessions, with both child's and mother's verbal interaction being a focus of analysis. This produced a total of roughly 1,250 quarter- to half-hour minute transcripts. Given traditional estimates of time required per transcript to compute multiple measures, we estimated a total time commitment of 6,250 hours to finish this part of the project, and the granting agency did not, in fact, predict that we would obtain any findings during the actual grant time window. However, *they were wrong*. This is because CLAN media linkage in Walker Controller, a CLAN program utility for transcription of spoken language, allows single keystroke playback of the segment being transcribed. This cuts down the time required to make an accurate transcript of the child's sample by roughly 75%. Moreover, because the transcriber can easily repeatedly compare the transcription to the original, accuracy is increased.

Next, we used the automated MOR function to assign and disambiguate grammatical descriptions of all the words in these 1,250 transcripts. The command "mor *.cha" will run MOR, POST, and MEGRASP in sequence on all target transcript files. The output has the form of this excerpt:

```
*CHI:mommy this xxx .
%mor:n|mommy pro:dem|this .
*CHI:these shoes on .
%mor:pro:dem|these n|shoe-PL adv|on .
*MOT:okay I can get her shoes on .
%mor:adj|okay pro:sub|I mod|can v|get det:poss|her n|shoe-PL adv|on .
*CHI:+< tiger .
%mor:n|tiger .
*MOT:is that a tiger ?
%mor:cop|be&3S pro:rel|that det:art|a n|tiger ?
*MOT:or is that a zebra ?
%mor:coord|or cop|be&3S pro:rel|that det:art|a n|zebra ?
*CHI:zebra .
%mor:n|zebra .
```

Following the running of MOR and POST, we then used the KIDEVAL command to generate spreadsheet output of each child's (and parent's) language features on more than two dozen variables. Some of these variables, such as pause length and MLU, are common across languages; others involving specific morphological features are unique and configurable to each language.

What about Norms?

In reviewing the literature on clinical use of language samples, LSA appears to be used most often when standardized test data cannot be obtained or are difficult to interpret. It seems to be particularly favored for assessment of very young children. However, there are conceptual issues in LSA for children at 24 months of age, which was the outcome measurement period for the toddlers in our study. Many of the normative or reference values are based on relatively few cases at lowest age ranges. For example, for MLU, a relatively recent report (Rispoli, Hadley, and Holt 2008) included 37 children at 24 months. Miller and Chapman (1981), the classic reference for MLU in clinical practice, reported on only 16 children in this age bracket, while the largest recent study to report expected values for MLU (as well as number of different words, NDW) (Rice et al. 2010) had 17 typically developing and 6 late-talking participants in the age bracket from 2;6 to 2;11. These are not extremely large populations on which to generalize impressions of a child's linguistic profile, which is why some researchers have expressed serious concerns about using MLU to identify whether a child is typically developing or impaired (Eisenberg, Fersko, and Lundgren 2001).

For Type-Token Ratio (TTR) or NDW, the situation is similar, since most of the studies referenced above also reported these measures, and few additional studies are available. For DSS and IPSYN, reference cohorts are similarly restricted. DSS reference tables report on only 10 children from 24 to 27 months of age (Lee 1974). In this age range, IPSyn provides data for 15 children (Scarborough 1990).

Our study does not intend to contribute normative data on these measures at this time. However, we can illustrate how the children in our study performed on these measures (all were typically developing, as is often the case in research reports taken from relatively high SES families). In general, data from this sample show values for MLU, DSS, and IPSYN that are consistent with prior, smaller samples (see figures 8.1–8.3).

These data suggest that KIDEVAL is a useful clinical tool for the assessment of spontaneous language data in 24-month-old children, a group for which few robust measures of LSA performance exist. Our results are comparable, and computed automatically, to data derived from much more time-intensive manual coding. However, we do note that the unaffected sample of Rice et al. did achieve higher MLU values than the other comparison cohorts.

We also computed correlations among LSA values and standardized test outcomes at 24 months of age. We obtained significant but weak correlations that probably justify larger studies of the available measures for toddlers and their construct validity. For instance, we correlated the children's MLU with IPSYN and DSS values; correlations were significant. This should not be surprising, since both IPSYN and DSS award points for various syntactic elements, and utterances with longer MLU values have greater opportunity to contain such features. However, it is perhaps surprising that the actual

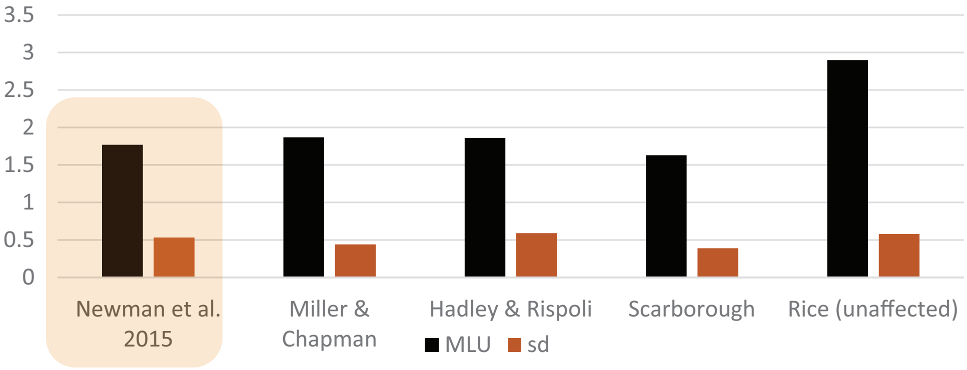


Figure 8.1
MLU values from prior research reports for children at 24 months of age. Note: Current = Newman et al., 2015, *n* = 122; Rice cohort is 2;6–2;11; combined *n* from other studies = 68. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

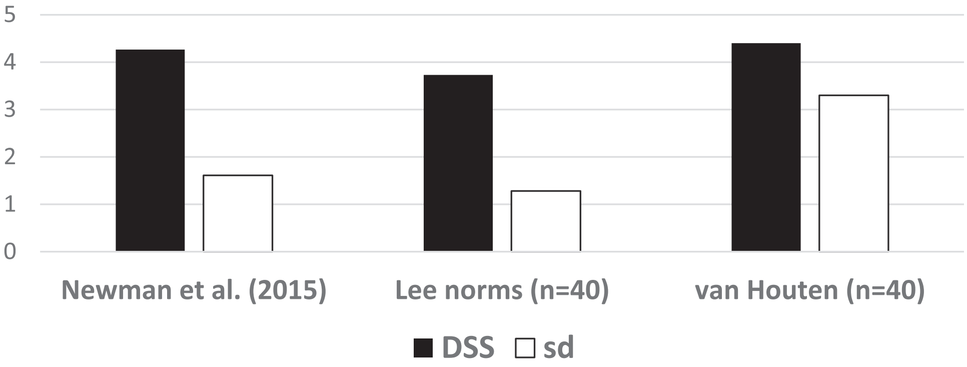


Figure 8.2
Developmental Sentence Score (DSS) values from Newman et al. (2015), reference values reported by Lee (1974), and values derived from the CHILDES van Houten corpus. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

correlations are relatively low, even though they reach significance given our large sample size. (See figures 8.4–8.6.) In particular, DSS correlates more poorly with MLU than does IPSYN, in all likelihood because fewer utterances at 24 months meet DSS eligibility standards and because very early utterances do not achieve DSS sentence points. Likewise, IPSYN and DSS do not correlate well with one another, probably for the same reasons, indicating that they are not interchangeable assessments of a toddler's language sample.

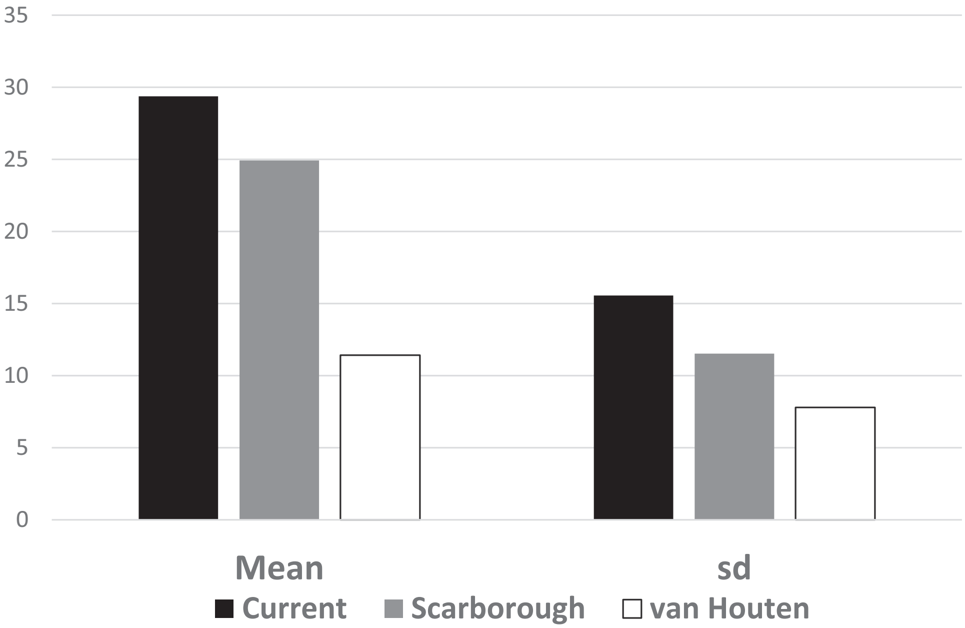


Figure 8.3
IPSYN values for Newman et al. (2015, "Current"), Scarborough (1990), and van Houten corpus (MacWhinney 1991). From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

Improving Norms

Our study suggests that, at young ages in English, some potential LSA measures do not appear to be measuring the same constructs. Clearly, a single LSA measure (especially MLU, which has been critiqued extensively; Eisenberg, Fersko, and Lundgren 2001) cannot provide the whole picture, and doing multiple LSAs is much too time consuming, unless more researchers and therapists use computer-assisted analysis to generate data that are more responsive to these concerns. We are, however, encouraged by the fact that the data from our large sample of toddlers do resemble those in smaller reference study reports. We also believe that psychometric evaluation of confidence intervals around mean values will be necessary to improve the robustness of measures such as DSS and IPSYN for distinguishing between typical and atypical performance, even though we do have some data to inform this decision-making process.

Fuller Support for SLPs

We are current working to move the CHILDES Project Archive from a repository and resource for researchers to a dynamic source of reference data that can be used to assess

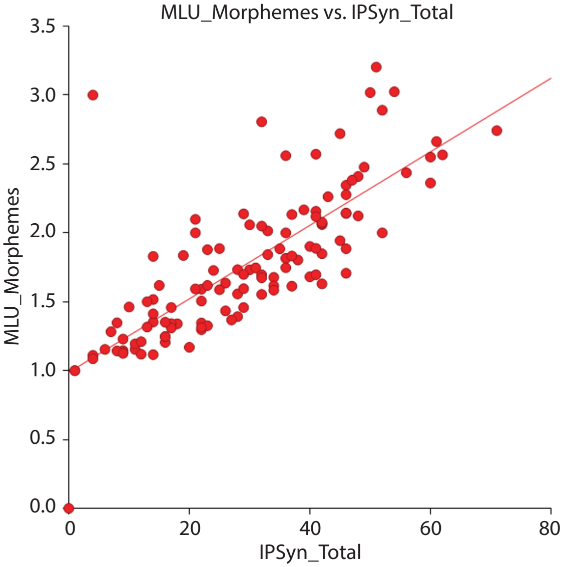


Figure 8.4
Correlation between MLU and IPSyn, r = .78, p = .000. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

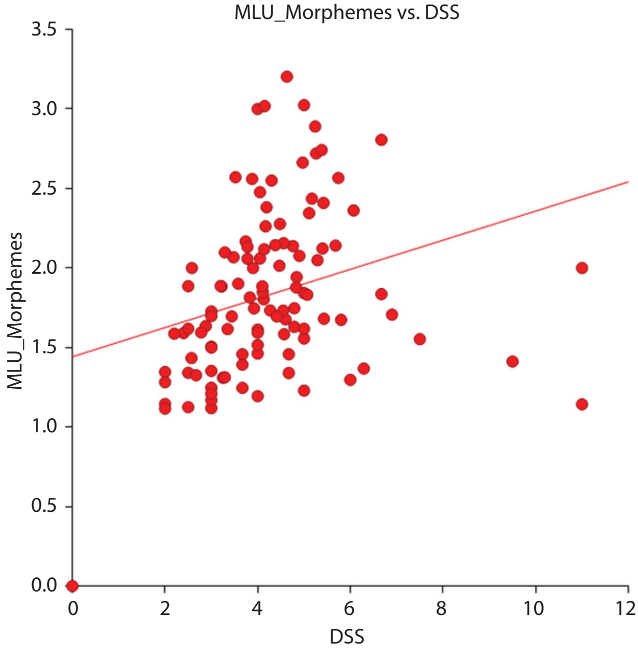


Figure 8.5
Correlation between MLU and DSS, r = .284, p = .003. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

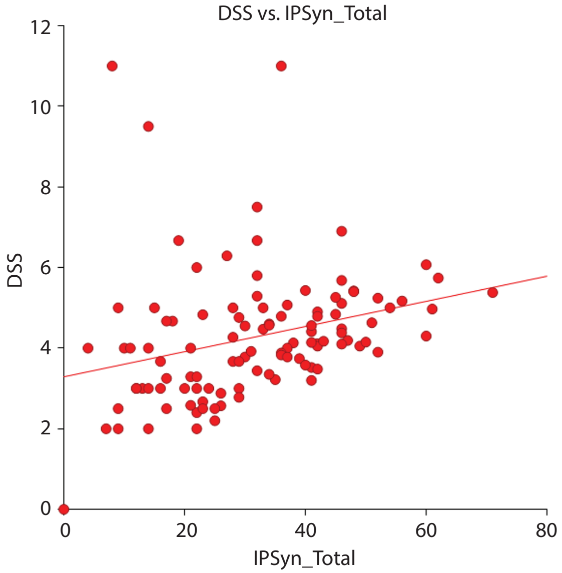


Figure 8.6
Correlation between DSS and IPSyn, r = .283, p = .00. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

and treat children across the world's languages. To this end, the TalkBank project is working to take the following actions that should greatly enhance clinicians' abilities to apply LSA to a broader range of children more easily and insightfully:

1. Increase the number of languages that can be automatically parsed and reported using CLAN utilities. As other contributors to this volume note, the free CLAN utilities now have grammars for a large number of languages; this number is growing yearly. Thus, clinicians working in Spanish, French, German, Dutch, Mandarin, Cantonese and other frequently used languages now have resources to perform accurate LSA of languages other than English.
2. Deploy existing corpora in the CHILDES Archive to improve "norms" for commonly used LSA outcome measures.

We are currently in the process of completing this second ambitious task. Recently, we completed KIDEVAL analysis of a large set of corpora (*n* = 630 children), all of whom spoke North American English, and all of whom were engaged in free play with their parents (a similar context). Results have been fairly interesting, and we provide only a brief taste of our findings here. First, we are happy to note that Roger Brown's (1973) observation that MLU is most useful when the child is fairly young or up until the point

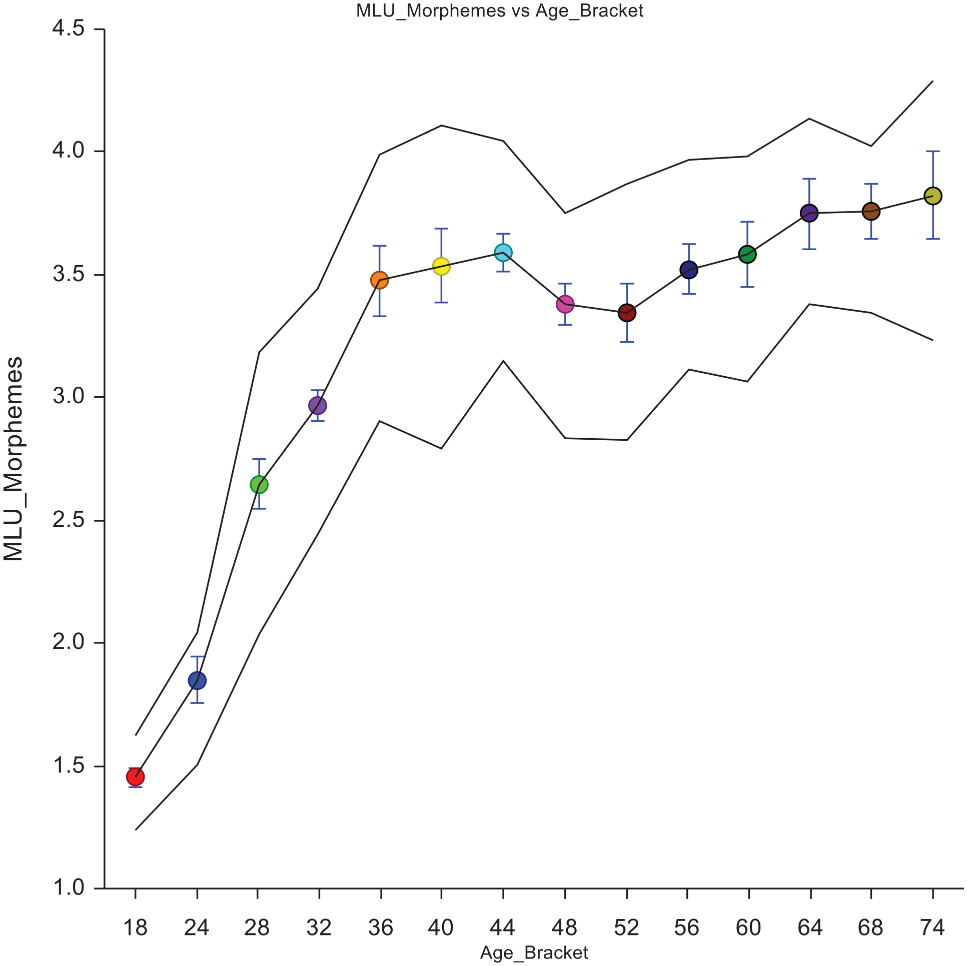


Figure 8.7

MLU values for 630 children in the CHILDES Archive. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

that it reaches a value of roughly 4.0 appears to be validated by this large sample, where MLU plateaus for our children past these values and ages (see figure 8.7).

We also note that IPSYN and DSS appear to be differentially sensitive to changes in age, as do two alternative ways of computing lexical (vocabulary) diversity—Type-Token Ratio (TTR) and vocd (Malvern et al. 2004), a computer algorithm less sensitive to variations in sample size. CLAN reports both in the KIDEVAL utility (see figures 8.8 and 8.9). Similar to our findings reported earlier for the Newman et al. study children, IPSYN and

DSS appear to measure different things, particularly across the broader age span covered by the CHILDES data. For example, IPSYN appears more sensitive to growth across very early childhood, whereas DSS appears to be more sensitive at older ages, perhaps as a function of the "sentence point" that provides more credit when a sentence is considered grammatical, an important construct in distinguishing typical from atypical development as children mature.

TTR and vocd (see figures 8.10 and 8.11) display a somewhat more difficult profile to evaluate. Vocd appears to track better with age across this sample than does TTR. Cur-

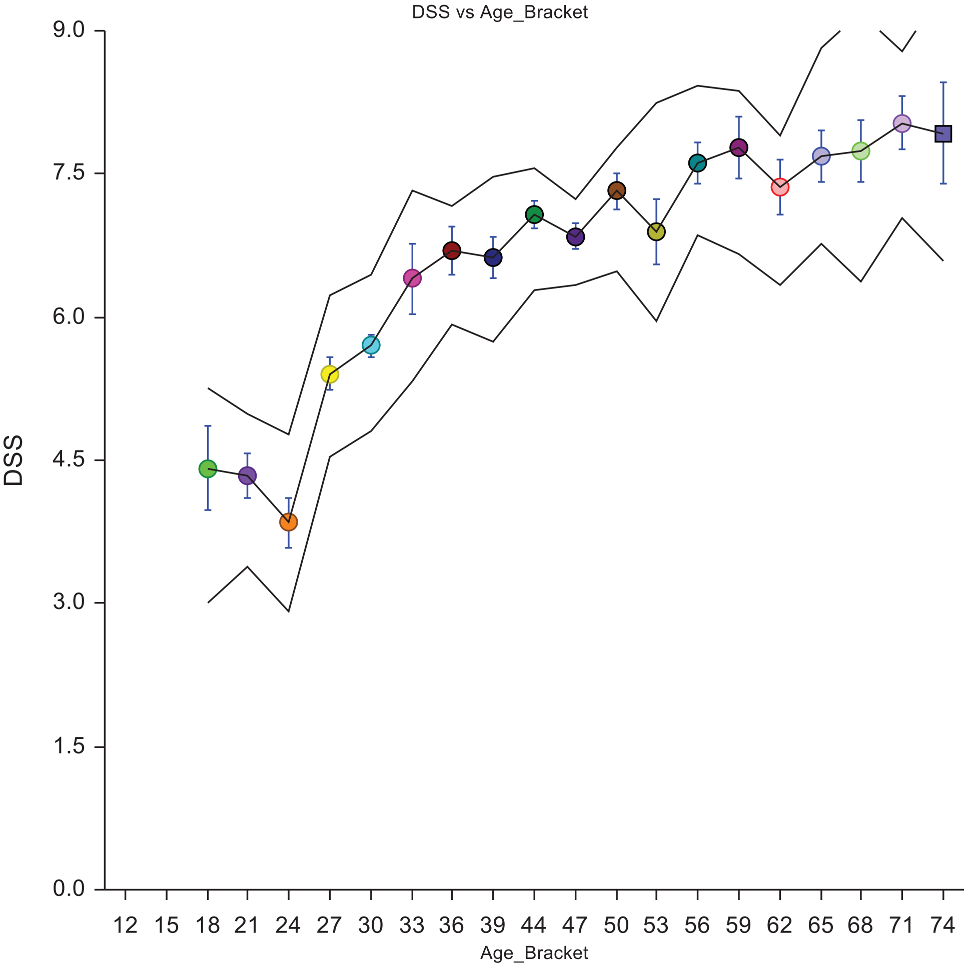


Figure 8.8
DSS scores for 630 children in the CHILDES Archive. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

rently, vocd is reported in a number of research reports (Pilar 2004; Silverman and Bernstein Ratner 2002; Owen and Leonard 2002; Wong 2010) but has no published norms; we hope to rectify this shortly. TTR has long been known to be vulnerable to a number of issues, particularly sample size; whether Vocd can improve on this to inform clinical assessment remains to be seen. Extending norms and evaluating the utility of various LSA measures is an ongoing initiative of great potential value to SLPs. We also note that there are no robust norms for LSA conducted with bilingual or English Language Learning (ELL) children, a major clinical cohort where LSA is used, given the parallel lack of standardized assessment norms for this population (Caesar and Kohler 2007).

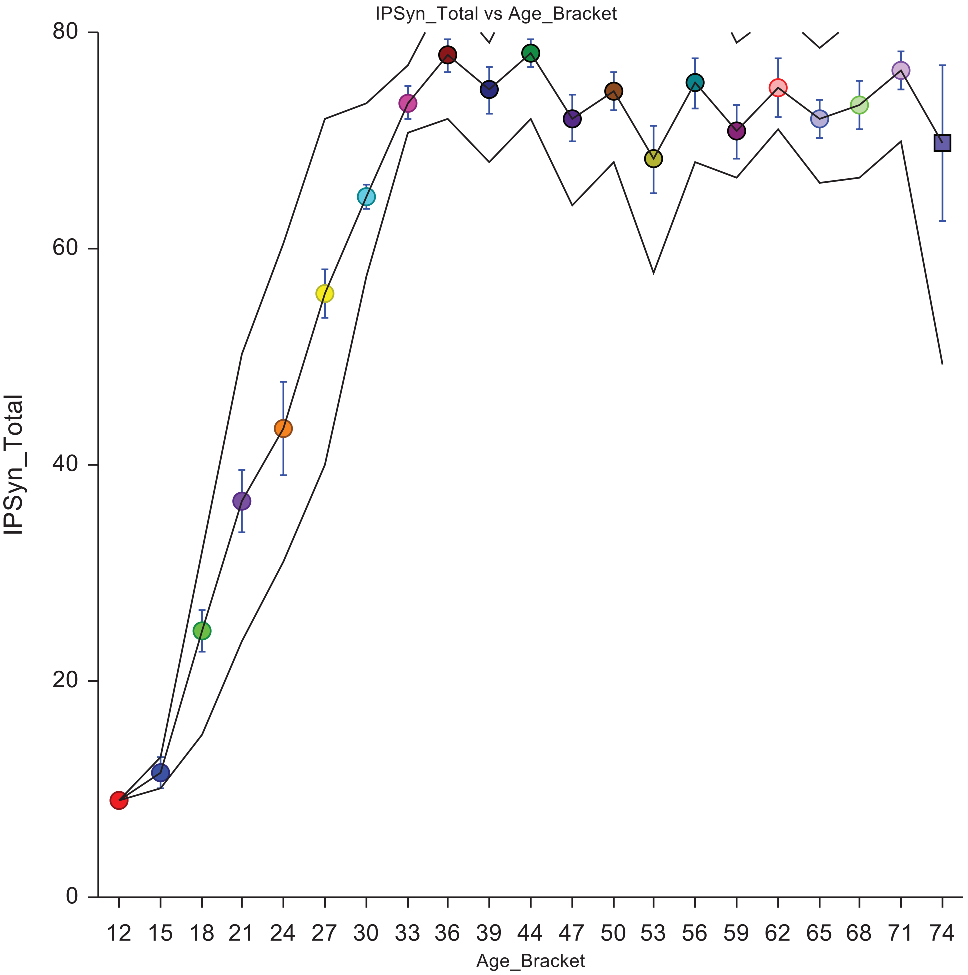


Figure 8.9
IPSyn scores for 630 children in the CHILDES Archive. From N. Bernstein Ratner and B. MacWhinney, “Your Laptop to the Rescue … ,” *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

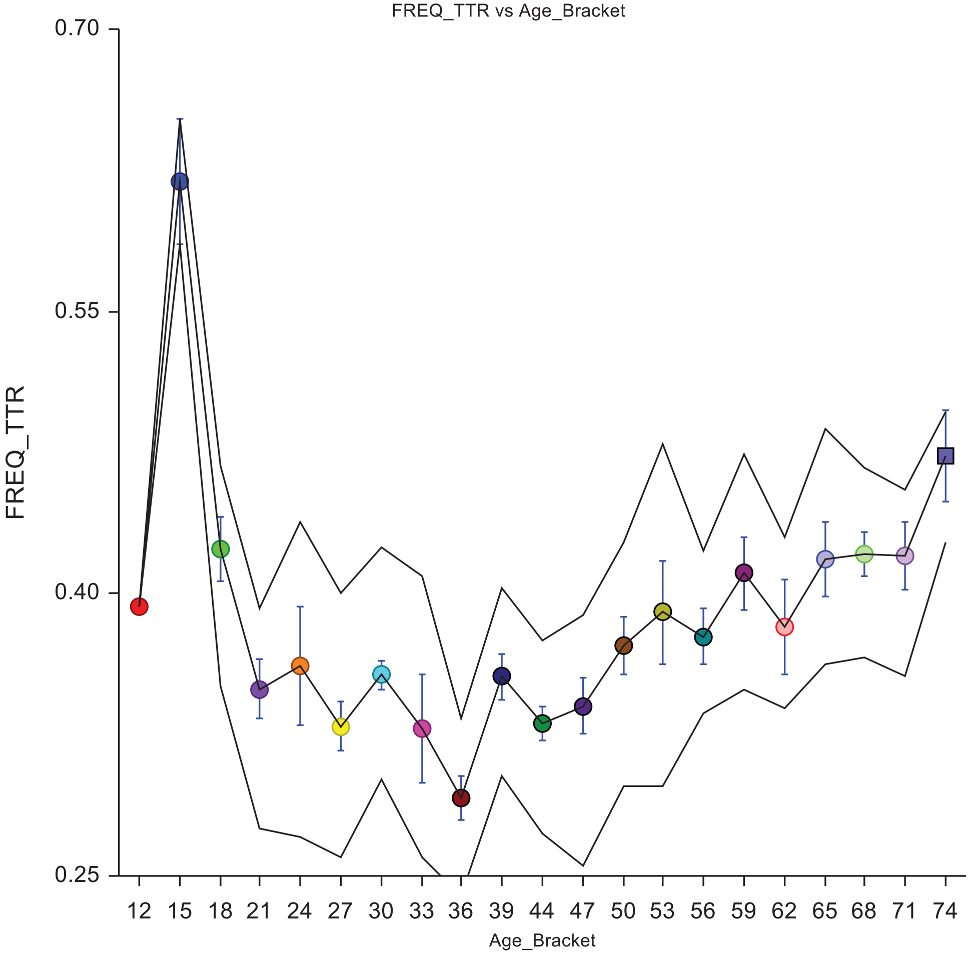


Figure 8.10
TTR values for 630 children in the CHILDES archive. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

Take-Away Messages

LSA is an important tool that one can use to appraise and understand child language ability in an ecologically valid way. Having said this, it is underutilized for a number of reasons, primarily because when done "by hand," it is very time-consuming. Because it is time-consuming, we know that clinicians do not fully exploit what can be learned from LSA, transcribing very short samples, and primarily deriving only a few measures such as MLU, which are not maximally informative for assessment, therapy planning, or outcome measurement. Media-linked transcription, such as is available using the free

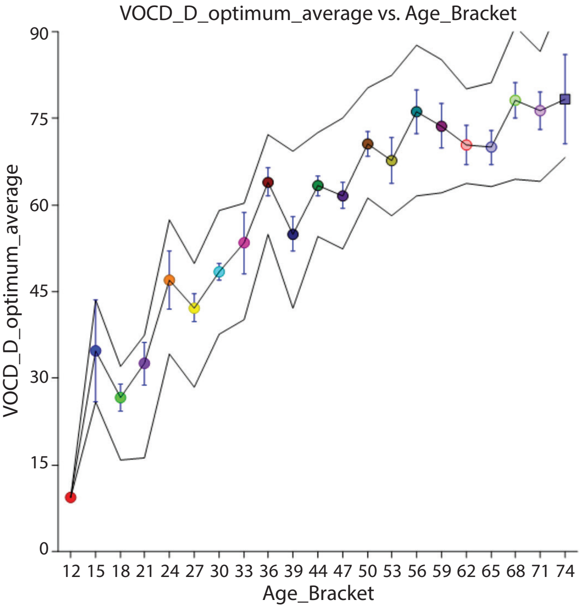


Figure 8.11
vocd values for 630 children in the CHILDES archive. From N. Bernstein Ratner and B. MacWhinney, "Your Laptop to the Rescue … ," *Seminars in Speech and Language* 37, no. 2 (2016): 74–84, www.thieme.com (reprinted by permission).

CLAN utilities available through TalkBank/CHILDES, greatly speeds transcription of a child's language sample. Once completed, this transcript can be used to generate many useful, accurately computed measures of child language performance. These can be used both to augment other assessment measures and to prioritize targets for intervention. Periodic LSA can also judge the child's progress in language growth, using the original LSA as a baseline measure. As clinically focused software evolves, the child's transcript can be paired with other utilities, such as PHON for phonological analysis, or FluCalc for fluency analysis, with little additional effort. CLAN grammatical parsers can also enable clinicians to evaluate bilingual children speaking a variety of languages, a unique benefit when working with a growing and challenging demographic in our profession.

When asked if they would use computer-assisted programs to analyze language samples more quickly and more informatively, the majority of clinicians in a recent survey agreed that they would, if they could identify how to accomplish this (Westerveld and Claessen 2014). We were intrigued to read of a successful pilot program to use SLP assistants or aides to generate transcripts and measures using SALT (Miller 2011), another LSA software program. Thus, we are optimistic that volumes such as this, along with web tutorials and the continued growth of programs available to SLPs, will help clinicians to exploit the potential of LSA more fully. In sum, the CHILDES/TalkBank utilities are an invaluable tool in an SLP's repertoire of clinical resources—free, time-saving, and computationally powerful. So power up your laptop and take computer-assisted LSA for a spin—for we predict that you will become a fast and loyal fan.

Broader Implications

We have examined in depth the ways in which the construction and validation of the KIDEVAL program rely on comparison of a given child language sample with the larger CHILDES database. A similar approach within the EVAL program enables us to compare a transcript from a given person who has aphasia with the fuller AphasiaBank database of 408 PWAs and 254 normal controls. Currently, we have only applied these methods for English and French, but they should work equally well for all 10 languages for which we can automatically compute morphosyntactic analyses.

We plan to build on our ability to automatically compute a wide variety of measures such as MLU, IPSyn, DSS, TTR, and 12 others, by developing norm-referenced clinical profiles such as KIDEVAL (for children) and EVAL (for adults with language impairment). Although a measure such as MLU involves a single construct, measures such as DSS, IPSyn, and QPA (Rochon et al. 2000) involve a complex combination of dozens of decisions about grammatical categories and errors. Using programs to automatically compute variant combinations of these underlying decisions, we will be able to learn which pieces of these larger scoring systems are most predictive of the actual level of language acquisition during development, using age as a proxy for developmental level. Work by Lubetich and Sagae (2014) has already shown that approaches based on data-mining methods such as classifier construction may be able to outperform these older standard measures. By gaining automatic access to large corpora that can be automatically analyzed, we will be able to test out these new and exciting possibilities for clinical diagnosis and developmental evaluation.

Acknowledgments

This work was supported by NSF grants BCS-1626294 and BCS-162-300, NIDCD grant DC008524, and NICHD grant HD082736 to Brian MacWhinney, and NIDCD grants DC015494 and DC017152, to Brian MacWhinney and Nan Bernstein Ratner, respectively.

References

Brown, Roger. 1973. *A First Language: The Early Stages*. Cambridge: Harvard.

Caesar, L. G., and P. D. Kohler. 2007. "The State of School-Based Bilingual Assessment: Actual Practice versus Recommended Guidelines." *Language, Speech and Hearing Services in Schools* 38:190–200.

Clahsen, H., and M. Rothweiler. 1992. "Inflectional Rules in Children's Grammars: Evidence from German Participles." In *Yearbook of Morphology*, edited by G. Booij and J. Van Marle. Dordrecht, Netherlands: Kluwer.

Cochran, P. S., and Julie J. Masterson. 1995. "Not Using a Computer in Language Assessment/Intervention in Defense of the Reluctant Clinician." *Language, Speech, and Hearing Services in Schools* 26:213–222.

Eisenberg, S. L., T. M. Fersko, and C. Lundgren. 2001. "The Use of MLU for Identifying Language Impairment in Preschool Children: A Review." *American Journal of Speech-Language Pathology* 10:323.

Evans, J. L., and J. Miller. 1999. "Language Sample Analysis in the 21st Century." *Seminars in Speech and Language* 20:101–198.

Finestack, L. H., and K. Satterlund. 2018. "Evaluation of an Explicit Intervention to Teach Novel Grammatical Forms to Children with Developmental Language Disorder." *Journal of Speech, Language, and Hearing Research* 61:2062–2075.

Freudenthal, D., J. Pine, and F. Gobet. 2010. "Explaining Quantitative Variation in the Rate of Optional Infinitive Errors across Languages: A Comparison of MOSAIC and the Variational Learning Model." *Journal of Child Language* 37:643–669.

Gorman, K. 2010. "Automated Morphological Analysis of Clinical Language Samples." Unpublished manuscript.

Hassanali, K.-N. 2014. "Automatic Generation of the Index of Productive Syntax for Child Language Transcripts." *Behavior Research Methods* 46:254–262.

Heilmann, John. 2010. "Myths and Realities of Language Sample Analysis." *Perspectives on Language Learning and Education* 17:4–8.

Hux, K. 1993. "Language Sampling Practices: A Survey of Nine States." *Language, Speech, and Hearing Services in Schools* 24:84–91.

Kemp, K., and T. Klee. 1997. "Clinical Language Sampling Practices: Results of a Survey of Speech-Language Pathologists in the United States." *Child Language Teaching and Therapy* 13:161–176.

Lee, L. 1974. *Developmental Sentence Analysis*. Evanston, IL: Northwestern University Press.

Lee, Laura, and Susan Canter. 1971. "Developmental Sentence Scoring: A Clinical Procedure for Estimating Syntactic Development in Children's Spontaneous Speech." *Journal of Speech and Hearing Disorders* 36:315–340.

Long, S., and R. Channell. 2001. "Accuracy of Four Language Analysis Procedures Performed Automatically." *American Journal of Speech-Language Pathology* 10:212–225.

Lubetich, Shannon, and Kenji Sagae. 2014. "Data-Driven Measurement of Child Language Development with Simple Syntactic Templates." COLING 2014, Dublin, Ireland.

MacWhinney, Brian. 1991. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Erlbaum.

MacWhinney, Brian. 2008. "Enriching CHILDES for Morphosyntactic Analysis." In *Trends in Corpus Research: Finding Structure in Data*, edited by H. Behrens, 165–198. Amsterdam: John Benjamins.

MacWhinney, Brian, and J. Leinbach. 1991. "Implementations Are Not Conceptualizations: Revising the Verb Learning Model." *Cognition* 29:121–157.

Malvern, David, Brian Richards, Ngoni Chipere, and Pilar Purán. 2004. *Lexical Diversity and Language Development*. New York: Palgrave Macmillan.

Marcus, Gary F., Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu, and Harald Clahsen. 1992. "Overregularization in Language Acquisition." *Monographs of the Society for Research in Child Development*, i–178.

Miller, J. F. 2001. "Focus on Schools: Having Trouble Monitoring Language Intervention? Language Sample Analysis is the Solution." *ASHA Leader* 6:5.

Miller, J. F. 2011. *Assessing Language Production Using SALT Software*. Middleton, WI: SALT Software.

Miller, Jon F., and Robin S. Chapman. 1981. “The Relation between Age and Mean Length of Utterance in Morphemes.” *Journal of Speech, Language, and Hearing Research* 24 (2):154–161.

Newman, Rochelle, Meredith Rowe, and Nan Bernstein Ratner. 2015. “Input and Uptake at 7 Months Predicts Toddler Vocabulary: The Role of Child-Directed Speech and Infant Processing Skills in Language Development.” *Journal of Child Language*, 1–16.

Overton, S., and Y. Wren. 2014. “Outcome Measurement Using Naturalistic Language Samples: A Feasibility Pilot Study Using Language Transcription Software and Speech and Language Therapy Assistants.” *Child Language Teaching and Therapy* 30:221–229.

Owen, A. J., and L. B. Leonard. 2002. “Lexical Diversity in the Spontaneous Speech of Children with Specific Language Impairment: Application of D.” *Journal of Speech, Language and Hearing Research* 45:927–937.

Pilar, D. N. 2004. “Developmental Trends in Lexical Diversity.” *Applied Linguistics* 25:220–242.

Pine, J. M., and E. V. M. Lieven. 1997. “Slot and Frame Patterns and the Development of the Determiner Category.” *Applied Psycholinguistics* 18:123–138.

Pinker, S., and A. Prince. 1988. “On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition.” *Cognition* 29:73–193.

Price, L. H., S. Hendricks, and C. Cook. 2010. “Incorporating Computer-Aided Language Sample Analysis into Clinical Practice.” *Language, Speech, and Hearing Services in Schools* 41:206–222.

Rice, Mabel, Filip Smolik, Denise Perpich, Travis Thompson, Nathan Rytting, and Megan Blossom. 2010. “Mean Length of Utterance Levels in 6-Month Intervals for Children 3 to 9 Years with and without Language Impairments.” *Journal of Speech, Language, and Hearing Research* 53:333–349.

Rispoli, Matthew, Pamela Hadley, and Janet Holt. 2008. “Stalls and Revisions: A Developmental Perspective on Sentence Production.” *Journal of Speech, Language and Hearing Research* 51 (4): 953–966.

Rochon, E., E. Saffran, R. Berndt, and M. Schwartz. 2000. “Quantitative Analysis of Aphasic Sentence Production: Further Development and New Data.” *Brain and Language* 72:193–218.

Scarborough, Hollis S. 1990. “Index of Productive Syntax.” *Applied Psycholinguistics* 11:1–22. doi: 10.1017/S0142716400008262.

Silverman, Stacy, and Nan Bernstein Ratner. 2002. “Measuring Lexical Diversity in Children Who Stutter: Application of vocd.” *Journal of Fluency Disorders* 27 (4): 289–304. doi: 10.1016/s0094–730x(02)00162–6.

Valian, Virginia, Stephanie Solt, and John Stewart. 2009. “Abstract Categories or Limited-Scope Formulae? The Case of Children’s Determiners.” *Journal of Child Language* 36 (04): 743–778.

Westerveld, M. F., and M. Claessen. 2014. “Clinician Survey of Language Sampling Practices in Australia.” *International Journal of Speech-Language Pathology* 16:242–249.

Wexler, K. 1998. “Very Early Parameter Setting and the Unique Checking Constraint: A New Explanation of the Optional Infinitive Stage.” *Lingua* 106:23–79.

Wong, Amy. 2010. “Differentiating Cantonese-Speaking Preschool Children with and without SLI Using MLU and Lexical Diversity (D).” *Journal of Speech, Language, and Hearing Research* 53:794–799.

9

Enabling New Collaboration and Research Capabilities in Language Sciences: Management of Language Acquisition Data and Metadata with the Data Transcription and Analysis Tool

María Blume, Antonio Pareja-Lora, Suzanne Flynn, Claire Foley,
Ted Caldwell, James Reidy, Jonathan Masci, and Barbara Lust

Introduction

The study of language is by definition interdisciplinary. It is situated at the intersection of the humanities and the social sciences. To investigate the human capacity for language knowledge, use, and acquisition, the field of linguistics must integrate scientific methods and situate itself among the approaches of the various other fields of cognitive science. Critically, fundamental questions—such as what it means to know a language or how a person acquires a language—depend on cross-linguistic investigation, which can illuminate both the possibilities and the constraints on human language. Empowered by cyberinfrastructure, language study in pursuit of these questions can begin to integrate across disciplines and across languages in a new way and can participate in the science revolution envisioned early by the National Science Foundation's Blue-Ribbon Advisory Panel on CyberInfrastructure (Atkins et al. 2003; see also Lave and Wenger 1991) and pursued subsequently (e.g., Berman and Brady 2005; NSF 2007; Borgman 2007, 2015; Abney 2011; G. King 2011; and T. H. King 2011).

Our digital and networked age now enables unprecedented opportunities for capturing language acquisition data, subsequently extracting stored data for analysis and interpretation, and supporting necessary collaborative scholarship. It also presents new challenges. In this chapter, we investigate those opportunities and we exemplify an approach to them through a case study—the construction of a Virtual Linguistic Lab (VLL)[1] and its cyberinfrastructure development of data capture and analysis tools.

We first review opportunities and challenges related to data quality and data complexity in the field of language acquisition.[2.] Next, we describe the infrastructure of principles and best practices that underpin the VLL. Then we introduce a cybertool central to the VLL, the Data Transcription and Analysis (DTA) tool,[3] intended to enable data storage, extraction, and analysis that lead to cross-linguistic discoveries and foster collaboration. We will argue that the tool, based on systematic metadata and data labeling, as well as on flexible linguistic annotations, facilitates collaborative research across projects, research labs, languages, and disciplines. We illustrate the use of the cybertool in cross-linguistic study of the first language acquisition of syntax, involving both experimentally derived and natural speech language data, in pursuit of current research questions. Finally, we consider the challenges for integrating the DTA tool and similar databases and tools with Linked Open Data (LOD) approaches in linguistics (see Chiarcos and Pareja-Lora, this volume, for an introduction of this movement). We explore the development of import/export functions in support of the interoperability necessary to achieving technically facile Linked Open Data, including exchange between various databases and ontologies.

This project is important because the more that linguists and other researchers interested in language look at the same set of data from different perspectives, the more we increase our knowledge of language and the stronger our evidence becomes through converging analyses (T. H. King 2011). Therefore, it is essential to design tools that make collaboration more effective, thus helping linguistics "to move forward on all fronts" (T. H. King 2011, 4). The importance of such interlinkage is underscored in the seminal Linked Data vision of Berners-Lee (2006):[4] Any data point becomes more powerful and useful when it can be linked with other data points; the challenge is to structure and make available this interlinkage (Chiarcos, Nordhoff, and Hellmann 2012). As has been suggested, "with a greater abundance of data than any individual or team can analyze, shared data enables mining and combining, and more eyes on the data than would otherwise be possible" (Borgman 2015, 10) yet at the same time "releasing data and making them usable are quite different matters" (2015, 13). For the researcher in language acquisition and in linguistics in general, Linked Data advances opportunities for data access and comparison. One could access, link, and analyze data from many different databases, projects, and systems, CHILDES[5] (MacWhinney 2000; Bernstein Ratner and MacWhinney, this volume), the datasets held by many individual researchers, those indicated by OLAC (see Simons and Bird, this volume), the Language Archive,[6] and the DTA tool database that we describe here, for example. This may be particularly relevant in areas where data is still scarce (cf. Blume et al., this volume).

Opportunities and Challenges in the Age of Digital Data

The VLL and the DTA tool are designed to enable us to address fundamental questions of cognitive science that inherently require collaborative exploration across projects and disciplines and that ultimately must involve cross-linguistic comparisons. For example, any search for universal properties of language acquisition—whether following hypotheses led by linguistic theory (e.g., generative theory), typology frameworks, or functional theories—requires access to and calibration of data from across languages, which in turn involves collaboration among scholars and research groups. Any search for neural foundations of language acquisition requires data access across disciplinary boundaries for collaborative teams of biologists, neuroscientists, linguists, and psychologists. Comparisons across first- and second-language acquisition (and beyond), and/or across language impairment, require a structured comparison of data and developmental observations. Although all linguistics endeavors looking for universal linguistic properties and the foundations of language require some degree of collaboration, this is especially important for research on child language acquisition, since recording, transcribing, and coding child language data is both complex and time-consuming (Blume and Lust 2017).

Although new technologies offer great promise in collaborative work, they also pose challenges. It is well known that different researchers develop and use different schemes for documenting and archiving their data, often for historical or pragmatic reasons. Challenges of documentation may also arise within a single project. Over the course of a project or a strand of research, the range of data that needs to be captured may evolve, often in unpredictable ways that are influenced by other developments in the researchers' own work or in the field.

Work in linguistics related to this challenge has been under way for some time. For example, E-MELD, the Electronic Metastructure for Endangered Languages Data,[7] is one current effort to address the need for digital-data-archiving standards informed by linguists.

The General Ontology for Linguistic Description (GOLD;[8] Farrar and Langendoen 2003; Simons et al. 2004; Cavar, Cavar, and Langendoen 2015; Langendoen, Fitzsimmons, and Kidder 2005; Langendoen this volume) is an effort to develop an ontology for linguistic description on the web that can maximize the usefulness of linked linguistic data made available to the wider community (see Bender and Langendoen 2010 for review). The Open Language Archives Community (OLAC)[9] seeks to advance best practices in data archiving and further to create a network of data repositories. The European Open Linguistics Working Group (OWLG)[10] cultivates Open Data sources in linguistics, including relevant ontologies (OntoLingAnnot's ontologies, Pareja-Lora and Aguado de Cea 2010; Pareja-Lora 2012a, 2012b, 2013).

However, to more fully address the challenge of structuring linkage, we need primary research tools with the power to calibrate metadata, thus allowing dissemination and access, and also able to reach deeply into linguistic analyses of language data so as to link fields across languages, datasets, and projects (see "The Data Transcription and Analysis Tool Empowers Discovery in Experimental Data" section later in this chapter).[11] If a certain level of standardization is attained, the language researcher can pursue the promise of automatic annotation (as achieved by CHILDES for morphosyntactic annotation in as many as 10 languages now, cf. Bernstein Ratner and MacWhinney, this volume), which would greatly assist the data-creation process.

Cross-linguistic research also requires capture of precise information about different levels of linguistic representation (e.g., specific speech sounds in an utterance, morphological markings, the ways words are assembled in phrases and sentences). Recent technology enables the integration of many levels of data description and analysis, because linkage among data points allows data to be entered and manipulated easily across levels.[12.]

Addressing these challenges in the field of language acquisition requires tools with standardized formats for data capture, but also with flexibility and room to evolve. Given that a central goal in cross-linguistic studies of language acquisition is to discover the similarities and differences in developmental patterns across languages, it is particularly important not only to standardize data but also to assimilate facts derived from independently designed studies, attempting to trace patterns and draw conclusions from widely varying data collected in widely different ways (e.g., Phillips 1995 and Dye et al. 2004).[13] The complexity of language data extends beyond the characteristics of linguistic forms. It includes methodological and research design information, information about data provenance (metadata), and multimedia representation of data in addition to markup (i.e., coding) along multiple dimensions (e.g., linguistic properties of specific words, morphology, phrases, and sentences) during analyses (for a discussion of related metadata issues, see Lust et al. 2010).[14]

Capturing these many dimensions of data is time- and labor-intensive. Taking advantage of technological opportunities to capture numerous dimensions of data may lead to cumbersome and even counterproductive machinery if the technological tools are not designed to maximize efficiency of data capture and analysis that facilitates collaboration. Leveraging a structured digital environment not only enables the capture and extraction of multiple levels of critical linguistic information, but also benefits researchers widely.

First, metadata may be captured more systematically. Tools for data capture can prompt the researcher to include information on crucial metadata fields. Enhancements of this type help standardize metadata documentation across projects and laboratories and therefore support comparability (Lust et al. 2010; Blume and Lust 2017).

Second, data may be better preserved and accessed. If data are not preserved along with metadata validating data provenance, they cannot reliably sustain collaborative research, replication, or reanalysis. Preservation, however, brings its own challenges. As Bird and Simons (2003) and participants in the GOLD project have noted, as technology changes, data gathered in particular formats may risk loss, highlighting the need for a sustainable and reliable cyberinfrastructure.

Finally, good data capture systems can be used as educational tools in support of teaching data management skills. Students in all fields of language acquisition require extensive training not only in linguistic analysis of language data, but also on the observational and experimental methods used first to collect language data and then to establish the metadata necessary for their preservation, dissemination, and collaborative use. As pointed out by G. King:

> More importantly, when we teach we should explain that data sharing and replication is an integral part of the scientific process. Students need to understand that one of the biggest contributions they or anyone is likely to be able to make is through data sharing. (G. King 2011, 270)

In this chapter we will illustrate steps we have made toward creating a cybertool that is intended to enable efficient capture and preservation of language data in support of collaborative cross-linguistic language acquisition research, thereby linking more-global metadata annotation to more-specific linguistic data annotation.

Development of Web-Based Cyberinfrastructure in Support of the Language Sciences: The Virtual Linguistic Lab (VLL) Case Study and the DTA Tool

Development of cyberinfrastructure must involve not only technology development but also vested community development (see Borgman 2015 and Blume and Lust 2017, chapter 14, for a discussion of these issues). The Virtual Linguistic Lab (VLL) is the result of a project generated by founding members of a burgeoning Virtual Center for the Study of Language Acquisition,[15] whose goal is the creation of a cyber-enabled international and interdisciplinary virtual research and learning environment. It was developed to enable researchers to share calibrated research methods during the primary research process and also to practice and teach scientific methods and best practices for data collection and management in primary research. The VLL houses a series of web-based courses integrating synchronous and asynchronous forms of interactive information distribution that teach students specific procedures for investigating language knowledge. These courses are meant to be taught in conjunction with the VLL research methods manual (Blume and Lust 2017).

The DTA Tool

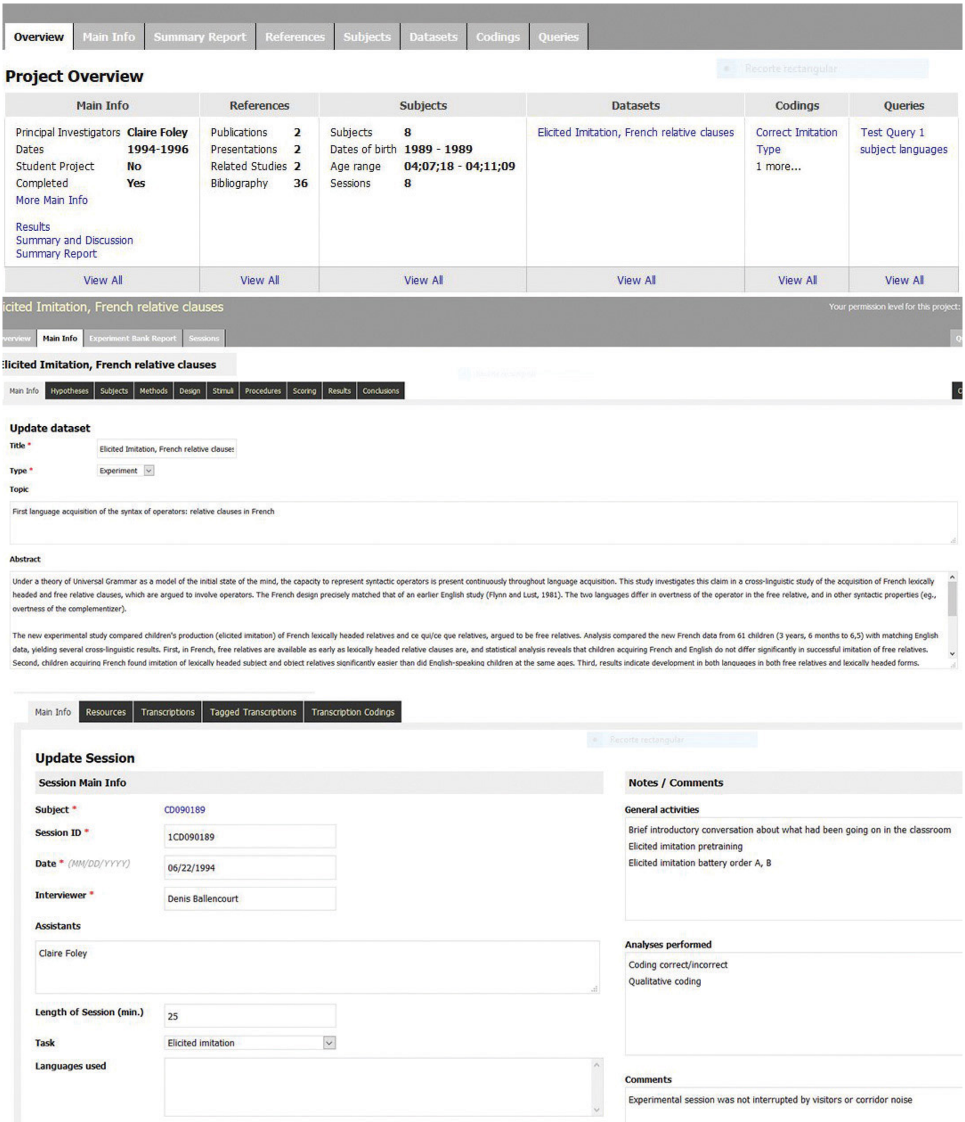
A web-based Data Transcription and Analysis (DTA) tool is part of the core of the VLL and its courses. The DTA tool provides a structured interface for metadata and data collection; it not only guides researchers and students in the primary research process, including data management, but it also results in a web-based calibrated database of continually expanding cross-linguistic data plus an Experiment Bank. The Experiment Bank records design and methodological factors connected with each particular experiment (or naturalistic study) through which language data are collected. The DTA tool follows the research principles and practices described in Blume and Lust (2017) and assumes that the researcher is familiar with them. The tool links to an associated set of continually expanding data from more than 20 languages collected over 30 years by the Cornell Language Acquisition Lab, other labs, and individual researchers across the United States and abroad. The DTA tool therefore provides a data bank resulting from the transcriptions and analyses it stores. However, it differs from other data banks, such as CHILDES, in that it is essentially designed as a primary research tool, structured to standardize metadata and data entry, management, and analysis; to permit the streamlined comparison of data across datasets and projects; and to foster sound collaborative research with shared data, as sketched below (see the VLL and VCLA websites for the VLL resources and for the vision and mission underlying the project; also Lust et al. 2005, 2010; Blume, Flynn, and Lust 2012; Blume and Lust 2012b, 2017, especially chapter 14).

The DTA tool provides a web interface that guides the researcher step by step through the processes of generating, storing, analyzing, and accessing data. It organizes data into projects that contain main information such as researcher names, purpose and leading hypotheses, results and discussion of the project, all to provide an overview of what the project is about. The project level also includes information on project participants (*subjects*), and references. Each project has complied with its institutions' IRB/Human Subjects criteria for approval. Intellectual property rights are protected by author (principal investigators) agreements. Human subjects' confidentiality is protected by allowing access to the full set of subject information only to authorized researchers (others can access the data, though confidential information is hidden). Each project has one or more datasets. The datasets are groups of data organized by any criteria relevant to the study (e.g., subject age, language, specific research task used), and include information on recording sessions, transcripts, and coding. The DTA tool guides users in the capture of information at the session level in four basic areas: *main information, resources, transcriptions,* and *codings*. The data fields on the session *main information* screen include metadata (e.g., duration and location of a session) that help to establish data provenance. A *resources* screen provides linkage to original, raw audio or raw video files, or to handwritten and scanned transcripts as well as field notes that provide data authenticity. Figure 9.1 exemplifies the project, dataset, and session levels of DTA tool coding.

Besides structuring the way that users contribute both data and a wide array of metadata (Pareja-Lora, Blume, and Lust 2013), the DTA tool allows one to link fields across datasets and projects. Through a password-protection system, individual users can be given access to individual projects or sets of projects. Project information (e.g., leading hypotheses, methodology, experimental batteries, or detailed subject information), results, and discussion can be added. The DTA tool tracks publications, related studies, and bibliography related to a research project. Thus, each project includes an experiment bank in which all the data, metadata, and aspects of a study are explained in detail. This level of detail allows researchers and students to know exactly how a particular study was conducted, and therefore it is fundamental to permit replication. Since replication is an increasingly important concern in social science research, knowing exactly how data were collected and analyzed is prerequisite for data reanalysis and for the creation of cross-linguistic comparative designs.

Research and Teaching

In addition to the DTA tool, course materials in the VLL include structured audiovisual materials for demonstration and practice; virtual workshops; a technical user's manual (Blume and Lust 2012a) to provide training for students on the use of the DTA tool and the Experiment Bank; teaching materials such as lecture slides; access to the research methods manual (Blume and Lust 2017); a set of materials to assist in data collection, data management, and data analyses (e.g., a multilingualism questionnaire for assessment of



Project Overview

Main Info	References	Subjects	Datasets	Codings	Queries
Principal Investigators **Claire Foley** Dates **1994-1996** Student Project **No** Completed **Yes** More Main Info Results Summary and Discussion Summary Report	Publications **2** Presentations **2** Related Studies **2** Bibliography **36**	Subjects **8** Dates of birth **1989 - 1989** Age range **04;07;18 - 04;11;09** Sessions **8**	Elicited Imitation, French relative clauses	Correct Imitation Type 1 more...	Test Query 1 subject languages
View All	View All	View All	View All	View All	View All

icited Imitation, French relative clauses

Your permission level for this project:

licited Imitation, French relative clauses

Update dataset

Title * Elicited Imitation, French relative clauses

Type * Experiment

Topic

First language acquisition of the syntax of operators: relative clauses in French

Abstract

Under a theory of Universal Grammar as a model of the initial state of the mind, the capacity to represent syntactic operators is present continuously throughout language acquisition. This study investigates this claim in a cross-linguistic study of the acquisition of French lexically headed and free relative clauses, which are argued to involve operators. The French design precisely matched that of an earlier English study (Flynn and Lust, 1981). The two languages differ in overtness of the operator in the free relative, and in other syntactic properties (eg., overtness of the complementizer).

The new experimental study compared children's production (elicited imitation) of French lexically headed relatives and ce qui/ce que relatives, argued to be free relatives. Analysis compared the new French data from 61 children (3 years, 6 months to 6,5) with matching English data, yielding several cross-linguistic results. First, in French, free relatives are available as early as lexically headed relative clauses are, and statistical analysis reveals that children acquiring French and English do not differ significantly in successful imitation of free relatives. Second, children acquiring French found imitation of lexically headed subject and object relatives significantly easier than did English-speaking children at the same ages. Third, results indicate development in both languages in both free relatives and lexically headed forms.

Update Session

Session Main Info

Subject * CD090189

Session ID * 1CD090189

Date * (MM/DD/YYYY) 06/22/1994

Interviewer * Denis Ballencourt

Assistants

Claire Foley

Length of Session (min.) 25

Task Elicited imitation

Languages used

Notes / Comments

General activities

Brief introductory conversation about what had been going on in the classroom
Elicited imitation pretraining
Elicited imitation battery order A, B

Analyses performed

Coding correct/incorrect
Qualitative coding

Comments

Experimental session was not interrupted by visitors or corridor noise

Figure 9.1
Combination of three different DTA tool screens: Project overview, dataset structure, and session main information. First language acquisition of relative clauses in French-Foley project and Elicited Imitation, French relative clauses dataset.[16]

degree and nature of multilingualism,[17] Blume and Lust 2017, 238, fn7); and platforms for long-distance discussion and collaboration. These materials are integrated into a cyberinfrastructure to accommodate the high-availability needs of distance learning programs (Blume, Flynn, and Lust 2012; Blume and Lust 2012b).

In the knowledge community of the VLL, eight universities in the United States and one in Peru[18] provided the foundation for VLL development and expansion, both nationally and internationally, by contributing to and participating in a first series of interuniversity courses that have been conducted on the basis of its resources. Founding members also contributed and shared both teaching and research examples and materials, leading to a diverse set of audiovisual samples available to researchers, teachers, and students alike. A set of publications develops in detail the educational aspects (Blume and Lust 2012b; Blume et al. 2014), technical aspects (Lust et al. 2005; Blume and Lust 2012a; Blume, Flynn, and Lust 2012; Pareja-Lora, Blume, and Lust 2013), and conceptual aspects (Lust et al. 2010) of both the VLL and the DTA tool (Blume and Lust 2017, 264–267).

In what follows we will exemplify the use of the DTA tool in pursuit of active research questions in the field of language acquisition. Using both experimental and natural speech data, we will instantiate the development of DTA tool annotations and query functions to address the challenges of specific and flexible data markups that are necessary for cross-linguistic analyses. Comparisons of English–French and English–Spanish data will exemplify, with data collected by the VLL community.

An Example of a Research Challenge: The Acquisition of Relative Clauses

One research area that has been confronted by the VLL, with support from the DTA tool, involves the acquisition of relative clauses. The complexities of this area require integrating specific data capture at the sentential level with the metadata represented in the DTA tool (e.g., figure 9.1). For example, relative clauses involve essential properties of natural language grammars, such as embedding of clauses within larger syntactic units, and also involve the structure of elements within an embedded clause. For example, in example (1) below, the relative clause *[that Natalia wrote]* is embedded within the object of the main clause under the noun head, *the book*.

(1) Max read [**the book** [that Natalia wrote]]

The linguistic properties of relative clauses are manifested in different ways across languages—for example, relative clause headedness and the elements that introduce the relative clauses vary, as does their sentential position. (See Lust, Foley, and Dye 2015, for a review; cf. Flynn and Foley 2004; and Flynn et al. 2005 forthcoming.)

The nature and complexity of relative clauses lead to critical questions about their acquisition, including:

1. Does a similar developmental pattern of acquisition in relative clause structures appear across languages? Do some relative clause types universally emerge sooner than others?
2. What explains cross-linguistic similarities and differences in developmental patterns? Which universal principles or parameters may underlie the acquisition of these structures biologically, and what must be learned?

Answering these questions requires the capacity to both capture and later extract morphological and syntactic information gathered from language acquisition data across languages. Thus, it requires a markup capacity that is uniform enough to permit cross-linguistic comparisons but differentiated enough to capture cross-linguistic differences. A series of experiments across languages has addressed these questions. In this section, we will illustrate the complexities of markup that were required in analyzing data in this series of experiments.

In English, using an elicited imitation (EI) task and experimental design, Flynn and Lust (1980) studied children's production of lexically headed and headless relative clauses[19] (e.g., examples (2) and (3) below, respectively); data and metadata were entered in the DTA tool and archived for analysis, resulting in current reanalysis and new data comparisons (Lust et al. 2015; Flynn et al. forthcoming).

In EI tasks, participants imitate utterances that vary structurally by experimental design, under controlled administrative conditions, so that language data can be analyzed specifically with regard to designed hypotheses (see Lust, Flynn, and Foley 1996; Blume and Lust 2017, chapters 4–6). Such imitation behavior requires the subject to analyze and reconstruct stimulus sentence structure, including both meaning and form. Given sufficiently taxing utterance length, participants will fully imitate (reproduce the stimulus sentence) correctly (without deformation) those structures that their developing grammars can generate (Lust, Flynn, and Foley 1996). Therefore, in an EI task, critical data include both correct imitations, according to design factors, and also the type of changes that participants may make in their possible deformation of the target utterance. Thus, the data that need transcription and analysis in this experiment require complex markup of the language produced by the subject.

Consider first the English, lexically headed relative clause in example (2).

(2) Experimental stimulus (lexically headed relative clause)
Big Bird pushes **the balloon which bumps** Ernie.

(3) Child reformation (headless relative)
Big Bird pushes **what bumps** Ernie.

Markup of child utterances in this experiment must capture not only the headedness of the sentence in the elicited language production, but also whether a *wh*-form introduces the relative clause, along with other properties that may be relevant to the researcher's hypotheses.[20] Example (3) shows a frequent conversion that children made from a lexically headed relative such as the one on example (2) to a headless relative, with accompanying change in the *wh*-form. Capturing this conversion requires adequate markup to reflect not only the change in *wh*-form, which relates to underlying structural differences, but also the change in structure. It thus bears on the nature of the knowledge that underlies development over time.[21]

In a replication of this English experiment with monolingual French-speaking children, Foley (1996) also found that children often converted a lexically headed French structure, like that in example (4) below, to a headless relative structure, like that in example (5). Data were again entered into the DTA tool.

(4) Experimental stimulus (lexically headed relative clause)
Aladdin choisit **la chose que** Fifi achète.
Aladdin choose-3SG the.FEM thing that Fifi buy-3SG
'Aladdin chooses the thing that Fifi buys.'

(5) Child reformation (headless relative) (age 4;2)
Aladdin choisit **ce que** Fifi achète.
Aladdin choose-3SG *ce* that Fifi buy-3SG
'Aladdin chooses what Fifi buys.'

Like English, French requires markup that can capture the form introducing the relative clause—*que* and *ce que*—a markup differing from the one required for English. For example, the form *que* introducing a relative clause with a gap in the object position would be replaced in adult language by *qui* in a relative clause with a gap in the subject position. Table 9.1 summarizes cross-linguistic differences in the elements introducing these types of relative clauses in both English and French.

Table 9.1
Structure of elements introducing relative clauses

	Lexically headed relative clause	Headless relative
English	… [CP *which* [C ∅ […	… [CP *what* C ∅ […
French	… [CP ∅ [C *qui/que* […	*ce* [CP ∅ [C *qui/que* […

A language-specific markup capacity must capture the variation shown in table 9.1 as well as the full range of syntactic and semantic factors reflected in the structures of examples (4) and (5). Such markup specificity, in conjunction with shared project design, is required to discover commonalities and differences between the English and the French acquisition facts.

In sum, a cybertool that permits precise and relevant cross-linguistic comparisons across the structures in examples (2) through (5) must provide a way to extract and compare not only the basic metadata across subjects compared, but also discrete aspects of a child's linguistic utterance, including corresponding linguistic elements in related forms across languages. Cross-linguistic comparisons, necessary for pursuing fundamental questions regarding language acquisition, depend on rich markup capacity that not only will share certain dimensions across languages, allowing cross-linguistic comparability, but also will differ across languages, allowing for language-specific appropriate analyses. The markup capacity must involve individual data fields as well as relations among them. This motivates the development of a tool that offers a wide range of markup and coding options that are hypothesis-dependent but remain nevertheless standardized as much as possible, to permit principled comparisons across subjects and languages.

The Data Transcription and Analysis Tool Empowers Discovery in Experimental Data

In this section we illustrate several features of the DTA tool that have been developed in order to begin to precisely capture and extract linguistic information required by research on the acquisition of relative clauses.

Data Coding

The DTA tool enables the researcher to link data from project to project, from dataset to dataset, from language to language, through calibrated but flexible markup. This is established through researcher-established coding linked to research hypotheses. In its capacity for flexible data coding, over and above calibrated metadata coding, the DTA tool can leverage the accumulated wisdom of the community in a robust set of coding options for linguistic utterances. It thus enhances efficiency of data capture and comparison.

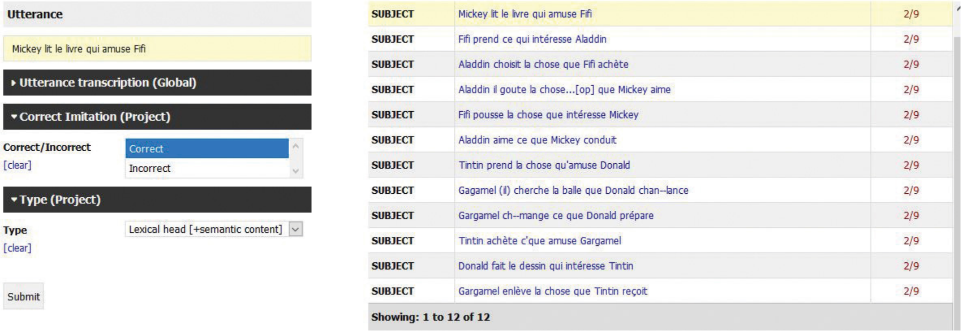
The DTA tool first establishes global codings—that is, metadata and data labels that are available across projects as standards.[22] All global codings are available for all researchers, but researchers can decide whether they want to use all of them or only a subset in each project. With the use of global codings across projects, all data can be calibrated, regardless of the specific question and subsequent hypothesis-specific coding of each project, since then some queries can be applied to all subjects. These codings allow the subsequent sorting and display of data that are focal to a research question, in a comparative way, thus significantly enhancing the interlinkage of data. In general, global codings in the tool can sort children by age, gender, MLU (mean length of utterance), and other basic properties of their productions in a session. The global codings were created and improved with the input of the members of the VCLA, underscoring the collaborative vision of the whole project.

In addition, more specifically, given a child's utterance in response to the Elicited Imitation relative-clause task, a basic coding marks up whether the language response is correct or not (following the standardized scoring criteria for the experiment). Another basic coding, given the hypotheses of the research design, assesses the type of headedness of the structure produced. Additional project-specific coding is then created to capture the complexities of each language. For example, French adds coding that specifies whether the relative clause is headed by *que* or *qui* and whether this would be the appropriate form in adult grammars.

For another example, the acquisition of relative clauses in Tulu (Somashekar 1999) requires additional language-specific coding for *correlative* and *verbal adjective* forms of relative clauses. These forms differ both in relative markers and in whether tense and agreement are marked on the relative clause verb. For all these studies, not only should markup permit the capture and glossing of all such cross-linguistic differences, but it must also permit coding of the range of changes that the child might or might not make to the stimulus verb, as well as any other changes accompanied by these conversions. Coding must also enable researchers to observe the cross-linguistic similarities and to compare child responses systematically across these three languages. For example, by such calibrated coding, the verbal adjective form in Tulu, which includes a null subject in the relative clause, and thus no overt internal clause head, has been discovered to surpass speed of development in either English or French headless relative clause acquisition (for discussion, see Somashekar 1999, chapter 8).

Once basic codes have been selected and/or created, the DTA tool allows the researcher to more specifically assess experimentally derived language data, examining one utterance at a time, using criteria established for the experiment and preparing it for cross-linguistic comparisons. Figure 9.2 illustrates this capacity.

In figure 9.2, the researcher has selected the highlighted utterance and applied the "correct/incorrect" global coding according to criteria standardized for the study. The DTA tool, which stores the experimental design and research criteria used for scoring of the specific project, includes a drop-down menu for project-specific coding (e.g., *Type*), allowing for both efficient data entry and reliability checking.



Utterance

Mickey lit le livre qui amuse Fifi

Correct/Incorrect: Correct

Incorrect

Type: Lexical head [+semantic content]

SUBJECT	Mickey lit le livre qui amuse Fifi	2/9
SUBJECT	Fifi prend ce qui intéresse Aladdin	2/9
SUBJECT	Aladdin choisit la chose que Fifi achète	2/9
SUBJECT	Aladdin il goute la chose...[op] que Mickey aime	2/9
SUBJECT	Fifi pousse la chose que intéresse Mickey	2/9
SUBJECT	Aladdin aime ce que Mickey conduit	2/9
SUBJECT	Tintin prend la chose qu'amuse Donald	2/9
SUBJECT	Gagamel (il) cherche la balle que Donald chan--lance	2/9
SUBJECT	Gargamel ch--mange ce que Donald prépare	2/9
SUBJECT	Tintin achète c'que amuse Gargamel	2/9
SUBJECT	Donald fait le dessin qui intéresse Tintin	2/9
SUBJECT	Gargamel enlève la chose que Tintin reçoit	2/9

Showing: 1 to 12 of 12

Figure 9.2
Application of project-specific codes in the DTA tool.

Queries

Once codes have been applied to the structured data, the DTA tool enables researchers to conduct queries. A query can produce a display of all utterances corresponding to a particular coding or set of codings, thus linking quantitative and qualitative data. In addition, the DTA tool allows the calculation of a mean number of correct or incorrect utterances for a particular sentence type (e.g., lexically headed relative clauses) in a particular experiment.[23] Then queries can link comparable data at various levels across one or more datasets and projects. The quantitative DTA output data allow integration with statistical analysis programs.

Cross-linguistic queries can be generated to test for commonalities and differences in development across languages as well as selected age ranges. For example, a query to calculate the mean number of correct productions (in this case, correct imitations of the target structures) for **all relative clause structures** in the age group 4;6 to 4;11 in French yields the result that 39% of the utterances were coded as correct. In contrast, for the matched dataset in English, the same query finds only 15% correct imitations. More refined queries, in terms of *Type* of relative coding reveal, however, that participants from the two languages produce correct responses for **headless relatives** at similar rates of success (approximately 50% are correct).

Thus, the analyses empowered by the DTA tool reveal the discovery that the rate of success for French children in fact matched that of English children for free relatives, although English lexically headed relative clauses took longer to emerge in target-like form than they did in French. This begins to identify where commonalities and differences lie in development across these two languages. Therefore, the power of the DTA tool consists not only in the systematic descriptions and computations it allows a user to perform, but also in its capacity to link and compare data across projects and datasets in a calibrated form.

Beyond revealing this quantitative trend, the DTA tool can also assist the researcher in integrating quantitative and qualitative data to help explain the trend, by pulling up transcriptions and relevant codings for each incorrectly imitated utterance, thus allowing researchers to qualitatively analyze the changes in children's language productions. For example, qualitative coding of children's changes on the model sentence reveals that children acquiring French frequently insert an overt operator in the headless relative structure, replacing *ce que* 'that' with *qu'est-ce que* 'what,' an overt operator that introduces questions (as in the example 5 and table 9.1 above). Because inclusion of this overt *wh*-form would not be predicted if the child were unaware of the structural position for such elements, the qualitative data provide additional evidence on the nature of developing knowledge—in this case, evidence that children are aware of the structural position of an overt *wh*-element even when the adult relative clause form does not fill that position and the children are computing this aspect of structure in their analyses of the relative clause structure. This French child conversion can be compared to the English child conversion in the section above ("An Example of a Research Challenge: The Acquisition of Relative Clauses"; see also Foley 1996; Flynn et al. forthcoming, for discussion of the theoretical significance of these results).

The DTA tool can help uncover similar knowledge through comparisons of responses within a language, for example across relative clause types. In Tulu, for instance, the correlative form includes overt marking of tense and agreement on the relative clause verb, in addition to a *wh*-form in the relativized position within the clause and a particular marker at the clause boundary. Somashekar (1999) found that when children converted the correlative form to the verbal adjective form, they not only changed the clause marker and omitted the *wh*-form, they also made required changes in verbal morphology, omitting tense and agreement, revealing their awareness of language-specific integration of syntactic elements.

In sum, use of the DTA tool aids discovery and theory development in several ways:

1. It can empower calibrated cross-linguistic comparisons. As an example, a query function displayed the fact that lexically headed relative clauses were imitated with less success in English than in French.
2. It can empower cross-linguistic comparisons in terms of coding specificity (e.g., relative clause as headed/headless). The query function revealed that English and French type "headless relatives" were imitated with similar percentages of "correct" forms, in spite of developmental differences between English and French.
3. It can empower explanation of descriptive data by linking both qualitative and quantitative data. In French, qualitative data analysis that is facilitated by the DTA tool suggests children's awareness of the CP structure in relative clauses, thus abetting the theory that CP structure variants across languages may explain developmental variations in language acquisition.
4. It can empower theory construction by cross-linguistic comparisons of calibrated transcription and coding. The markup capacity of the DTA tool permits close comparison of the cross-linguistic forms, highlighting a similarity across forms in French and English, as well as in Tulu, and suggesting a potentially universal developmental path for relative clause knowledge, as well as language-specific variation in specific relative clause forms (Flynn and Foley 2004; Flynn et al. 2005; Flynn et al., forthcoming).

The Data Transcription and Analysis Tool Empowers Discovery in Natural Speech Data

Analyzing Natural Speech Data

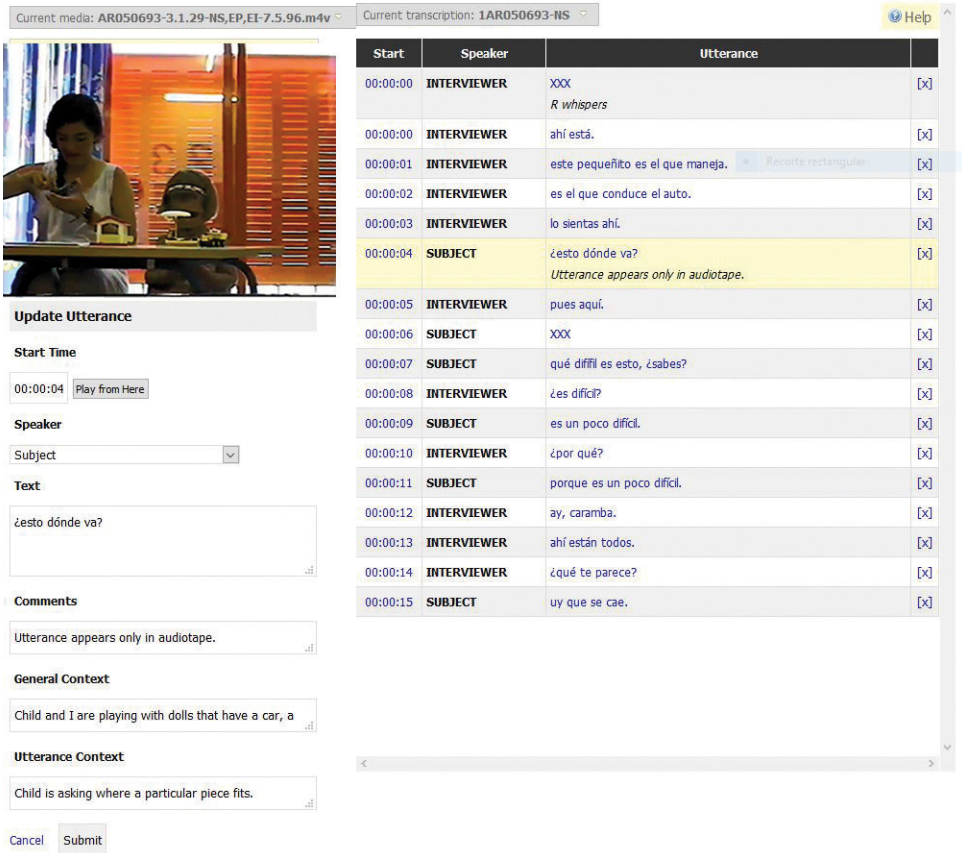
Natural speech data often complement experimental data in the study of language acquisition (Blume and Lust 2017). These data are less constrained than experimentally derived data, both in terms of types of language structures that the researcher may want to study and in terms of language productions that the speaker may generate. However, the DTA tool allows researchers to query across cross-linguistic projects and databases of natural speech systematically and then to search for utterances matching very specific characteristics that may be relevant to general as well as specific questions across projects and languages. For example, do both Spanish and English child language production patterns indicate similar rates of development in terms of length? More specifically, do they similarly display noninflected verbs? Theories have varied widely in terms of the significance of inflection omission in early child language. Debate on this issue requires consideration of the linguistic and pragmatic context of the verb utterance (Boser et al. 1992; Blume 2002; Dye 2011; and references cited within the three sources). Systematic cross-linguistic comparisons of child language through the DTA tool can inquire generally as to whether MLU development occurs at similar ages across languages.[24] They can also inquire as to whether children's use of noninflected verb forms occurs to the same degree and in the same contexts across languages.

Data Transcription

The DTA tool offers a structured process for reliable transcription of natural speech and experimental data, which is the first stage for data accession (cf. Blume and Lust 2017). Figure 9.3 shows a section of the DTA tool's transcription of a natural speech sample of a Spanish-speaking child from the Spanish Natural Speech Corpus–Blume project (Blume and Lust 2012a) for which both video and audio recordings are available. The transcription component enables each utterance to be tagged to a point in the video or audio (as shown in the *Start* column and the *Set start to* option), enabling access to information about the experimental context that may be important but potentially not recognized as such at the time of initial transcription. The tool offers a drop-down list for identifying a speaker (whether the subject being studied, the interviewer, or another speaker), a field for utterance entry, and a comments field.[25]

Coding Natural Speech

Here, based on calibrated global codings, we provide examples from two male subjects of the same age, who speak two different languages (Spanish or English), to illustrate how DTA tool codings can be applied in the pursuit of a research question for natural speech data. We show a set of codings established for this study below and also in figure 9.4. Each utterance can be tagged with 113 different global codings, subsets of which are exemplified in the figure. These codings were selected to guide new researchers to perform a basic description of the data (*Is this a sentence or not? If it is not a sentence, what sort of structure is it? What is the speech act intended by the utterance? Is it a simple or complex sentence?*). Further coding was created for a more detailed description of the major sentence functions and the structures they represent (subject, verb, direct and indirect objects; e.g., Radford 2004; Zagona 2001).



Current media: **AR050693-3.1.29-NS,EP,EI-7.5.96.m4v**

Current transcription: **1AR050693-NS**

Start	Speaker	Utterance	
00:00:00	**INTERVIEWER**	XXX *R whispers*	[x]
00:00:00	**INTERVIEWER**	ahí está.	[x]
00:00:01	**INTERVIEWER**	este pequeñito es el que maneja.	[x]
00:00:02	**INTERVIEWER**	es el que conduce el auto.	[x]
00:00:03	**INTERVIEWER**	lo sientas ahí.	[x]
00:00:04	**SUBJECT**	¿esto dónde va? *Utterance appears only in audiotape.*	[x]
00:00:05	**INTERVIEWER**	pues aquí.	[x]
00:00:06	**SUBJECT**	XXX	[x]
00:00:07	**SUBJECT**	qué diffil es esto, ¿sabes?	[x]
00:00:08	**INTERVIEWER**	¿es difícil?	[x]
00:00:09	**SUBJECT**	es un poco difícil.	[x]
00:00:10	**INTERVIEWER**	¿por qué?	[x]
00:00:11	**SUBJECT**	porque es un poco difícil.	[x]
00:00:12	**INTERVIEWER**	ay, caramba.	[x]
00:00:13	**INTERVIEWER**	ahí están todos.	[x]
00:00:14	**INTERVIEWER**	¿qué te parece?	[x]
00:00:15	**SUBJECT**	uy que se cae.	[x]

Update Utterance

Start Time
00:00:04

Speaker
Subject

Text
¿esto dónde va?

Comments
Utterance appears only in audiotape.

General Context
Child and I are playing with dolls that have a car, a

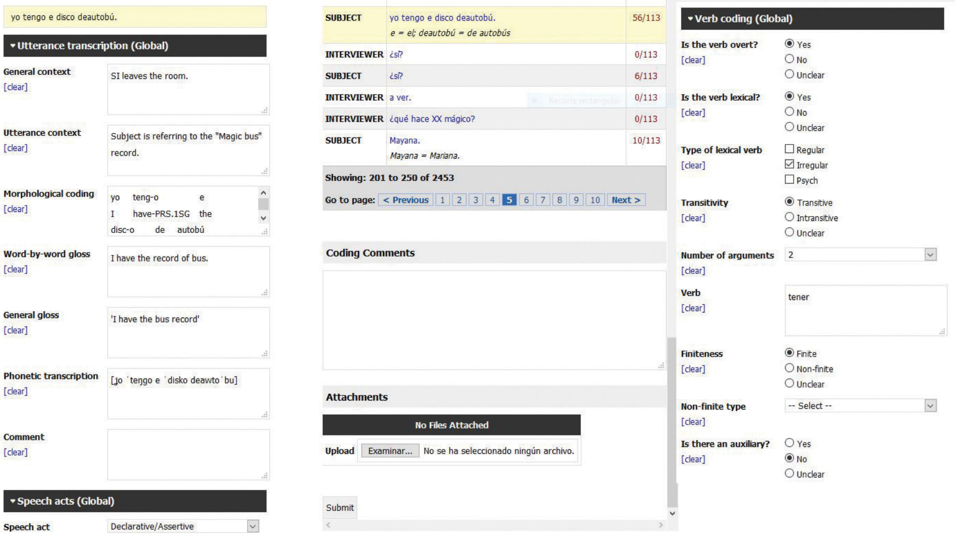
Utterance Context
Child is asking where a particular piece fits.

Figure 9.3
Transcription screen. Natural Speech Corpus–Blume project.

The first subject (04BG021097) was 2;02,00[26] at the time of recording (English Natural Speech Corpus–Lust; Blume and Lust 2012a). His data are compared here to those of 01RP071296, who was 2;01,28 at the time of recording (Spanish Natural Speech Corpus–Blume).

Figure 9.4 shows how two of the coding sets (*utterance transcription* and *verb*) are displayed on the coding screen, for the Spanish-speaking child.[27]

The contents of the first coding set applied, "utterance transcription coding," is exemplified in the upper part of figure 9.4. This coding set allows for the input of contextual/pragmatic information on the utterance and on the session information (i.e., *general context* and *utterance context*), and also for morphological coding,[28] glossing (word-by-word and a freer meaning-preserving gloss), and IPA transcription of the utterance itself.[29]



yo tengo e disco deautobú.

Utterance transcription (Global)

General context: SI leaves the room.

Utterance context: Subject is referring to the "Magic bus" record.

Morphological coding:

```
yo    teng-o         e
I     have-PRS.1SG   the
disc-o      de    autobú
```

Word-by-word gloss: I have the record of bus.

General gloss: 'I have the bus record'

Phonetic transcription: [ʝo ˈteŋgo e ˈdisko deawtoˈbu]

Comment:

Speech acts (Global)

Speech act: Declarative/Assertive

SUBJECT	yo tengo e disco deautobú. *e = el; deautobú = de autobús*	56/113
INTERVIEWER	¿si?	0/113
SUBJECT	¿si?	6/113
INTERVIEWER	a ver.	0/113
INTERVIEWER	¿qué hace XX mágico?	0/113
SUBJECT	Mayana. *Mayana = Mariana.*	10/113

Showing: 201 to 250 of 2453

Coding Comments

Attachments

No Files Attached

Upload: No se ha seleccionado ningún archivo.

Verb coding (Global)

Is the verb overt? ● Yes ○ No ○ Unclear

Is the verb lexical? ● Yes ○ No ○ Unclear

Type of lexical verb: ☐ Regular ☑ Irregular ☐ Psych

Transitivity: ● Transitive ○ Intransitive ○ Unclear

Number of arguments: 2

Verb: tener

Finiteness: ● Finite ○ Non-finite ○ Unclear

Non-finite type: -- Select --

Is there an auxiliary? ○ Yes ● No ○ Unclear

Figure 9.4
Upper section: Utterance transcription codings. Lower section: A part of the verb-coding set. Both applied to a Spanish-speaking child of the Spanish Natural Speech Corpus–Blume.

Of the six global coding sets chosen for this project, the first is *Speech Acts* and contains codes for the speech act intended by the utterance and discourse nature of the utterance (e.g., Is it spontaneous, or is it a question answer, or repetition?). The second set, *Basic Linguistic* coding, starts to inquire about the complexity and structure of the utterance (e.g., Is it a multi-word utterance? Is it a sentence? How many words, morphemes and syllables does it contain?). The third set, *Non-sentence*, allows the researcher to determine the structure of the utterance if it is not a sentence (e.g., Is it a noun phrase, an adjectival phrase, or a fragment?). The fourth coding set, *Clause type*, defines whether the utterance is a matrix or an embedded sentence, whether it contains an overt complementizer, and whether it is negated. Finally, the *Verb, Subject, Direct Object*, and *Indirect Object* coding sets characterize those specific structures and functions, if they actually appear in the utterance. The definitions for these codings can be found in Blume and Lust (2017) and are linked to the DTA tool experiment bank, thus empowering replication and reliability of coding. There were 113 codings in total (organized in six coding sets) that could be applied to each utterance.

A Case Study: Spanish–English Comparison of Tense and Finiteness

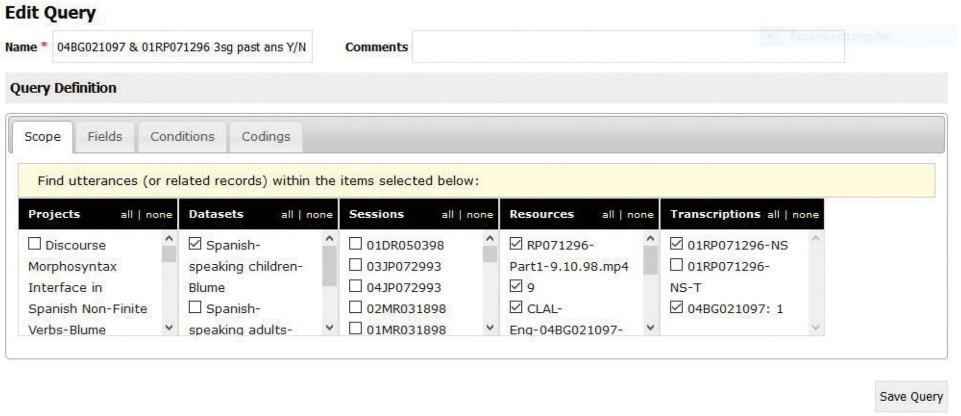
Once a set of basic global codings has been applied to the transcripts of subjects, both general and specific cross-language and cross-project queries can be run. A general query, for example, can be an MLU query.[30]

Although these two male subjects have similar ages, the MLU query revealed the Spanish-speaking child's MLU in morphemes (4.14) to be higher than that of the English-speaking child (2.29), although both were coded according to our calibrated MLU criteria (Blume and Lust 2017).

Other queries can isolate the data relevant to a specific hypothesis that in turn is related to a particular research question. We provide two examples here. The first query addresses a basic descriptive question related to the frequent lack of verbal inflection in child language: Do both matched Spanish and English child language samples indicate a similar proportion of production of noninflected verbs that would be grammatical for adult grammar over those that would not be (thus distinguishing between cases where either the pragmatic or the linguistic context allows noninflected verbs from those cases where the noninflected verb is not licensed this way, and therefore it is ungrammatical in adult language)?[31] More specifically, do they occur in both spontaneous and question/answer contexts? (For example, when one is asked, *What are you doing?*, it is perfectly fine to answer *Ø writing a paper* in adult English as well as in Spanish.) The second query demonstrates the DTA tool's capability; a researcher could, for example, also query for all utterances produced by children with a third person singular present tense verb in declarative sentences as a reply to a yes/no question.

For such queries we first set their scope, as shown in figure 9.5. In this screen researchers select which projects, datasets, resources, and transcriptions the query will apply to.

For this case study, under "projects" we have selected both projects "English Natural Speech Corpus-Lust" and "Spanish Natural Speech Corpus-Blume." The datasets column displays the available datasets for those projects. We selected "English-speaking children-Lust" and "Spanish-speaking children-Blume," as exemplified for Spanish in figure 9.5.



Edit Query

Name * 04BG021097 & 01RP071296 3sg past ans Y/N **Comments**

Query Definition

Find utterances (or related records) within the items selected below:

Projects all | none	Datasets all | none	Sessions all | none	Resources all | none	Transcriptions all | none
☐ Discourse Morphosyntax Interface in Spanish Non-Finite Verbs-Blume	☑ Spanish-speaking children-Blume ☐ Spanish-speaking adults-	☐ 01DR050398 ☐ 03JP072993 ☐ 04JP072993 ☐ 02MR031898 ☐ 01MR031898	☑ RP071296-Part1-9.10.98.mp4 ☑ 9 ☑ CLAL-Eng-04BG021097-	☑ 01RP071296-NS ☐ 01RP071296-NS-T ☑ 04BG021097: 1

Figure 9.5
Query scope.

Table 9.2
Query fields

Table/Field
Session: Title or Transcription: Title
Session: Age
Utterance: ID
Utterance: Speaker
Utterance: Text
Utterance: Comments

The session column then displays the sessions inside those datasets from which we selected the session titles “04BG021097” and “01RP071296”;[32] then we select the resources and transcripts for both Spanish and English sessions that we mean to compare. Next, we select the fields that we want the query to display. In this case, we select the fields shown in table 9.2.

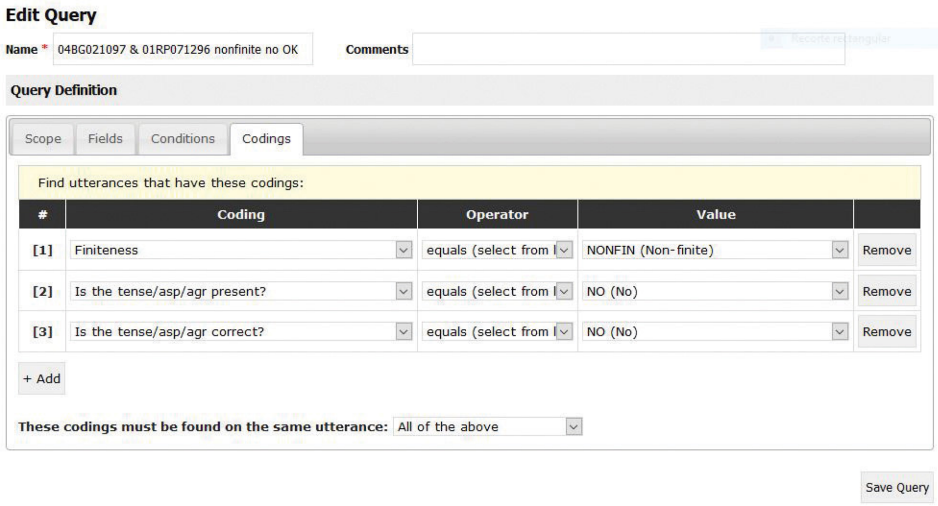
Under “Conditions” we include “Utterance: Speaker” equals “SUBJECT” to limit the query search to the utterances produced by the child—the subject being studied. The scope and conditions are the same for both queries.

In our first query, we look for sentences headed by noninflected verbs (infinitives, present or past participles, or bare verbs) used in contexts where they would be ungrammatical for adults. We next need to set the coding titles and coding values we want the query to search for under the tab *Codings*, as shown in figure 9.6.

For this query, we set coding titles and values to “finiteness equals non-finite,” “is the tense agreement present? equals no,” and “is the tense agreement correct? equals no.” We get the results shown in figure 9.7.

The DTA tool allowed us to process a total of 892 child utterances (across both Spanish and English samples), selecting three of those utterances that were critical to testing hypotheses regarding the development of verb inflection.

Here we see that the English-speaking child, 04BG021097, produced five utterances (20.8%), while the Spanish-speaking child, 01RP071296, produced four utterances (6.5%) that were sentences with noninflected verbs that were ungrammatical in adult language. In contrast, other queries revealed that, out of a total of 24 sentences, the English-speaking child produced 18 sentences with correct inflection (75%) and one sentence with a noninflected verb that was correct for the adult grammar (4.2%). Out of 62 sentences, the Spanish-speaking child, 01RP071296, produced 57 sentences with correct inflection (91.9%) and one sentence with a noninflected verb that was correct for the adult grammar (1.6%). Although we ran the query on only two subjects for demonstration purposes, the query results show a well-known pattern in which grammatical sentences outnumber ungrammatical ones, and in which verbs with inflection outnumber noninflected ones. We also find, in this particular case, that the Spanish-speaking child appears more linguistically



Edit Query

Name * 04BG021097 & 01RP071296 nonfinite no OK

Comments

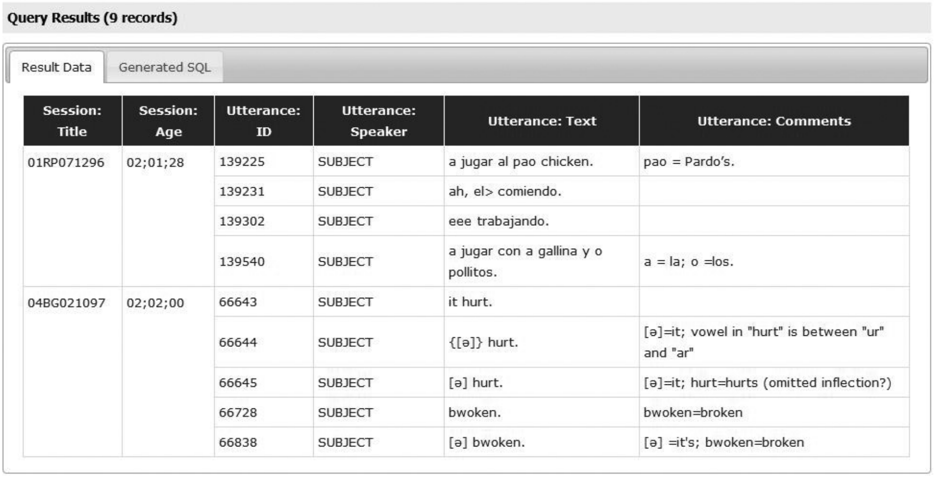
Query Definition

Find utterances that have these codings:

#	Coding	Operator	Value
[1]	Finiteness	equals (select from l	NONFIN (Non-finite)
[2]	Is the tense/asp/agr present?	equals (select from l	NO (No)
[3]	Is the tense/asp/agr correct?	equals (select from l	NO (No)

These codings must be found on the same utterance: All of the above

Figure 9.6
Sentences headed with noninflected verbs that are ungrammatical in adult language codings.



Query Results (9 records)

Session: Title	Session: Age	Utterance: ID	Utterance: Speaker	Utterance: Text	Utterance: Comments
01RP071296	02;01;28	139225	SUBJECT	a jugar al pao chicken.	pao = Pardo's.
		139231	SUBJECT	ah, el> comiendo.	
		139302	SUBJECT	eee trabajando.	
		139540	SUBJECT	a jugar con a gallina y o pollitos.	a = la; o =los.
04BG021097	02;02;00	66643	SUBJECT	it hurt.	
		66644	SUBJECT	{[ə]} hurt.	[ə]=it; vowel in "hurt" is between "ur" and "ar"
		66645	SUBJECT	[ə] hurt.	[ə]=it; hurt=hurts (omitted inflection?)
		66728	SUBJECT	bwoken.	bwoken=broken
		66838	SUBJECT	[ə] bwoken.	[ə] =it's; bwoken=broken

Figure 9.7
Sentences headed with noninflected verbs that are ungrammatical in adult language results.

advanced than the English-speaking one, since he produces more sentences, and the proportion of grammatical versus ungrammatical sentences is also higher in his case.

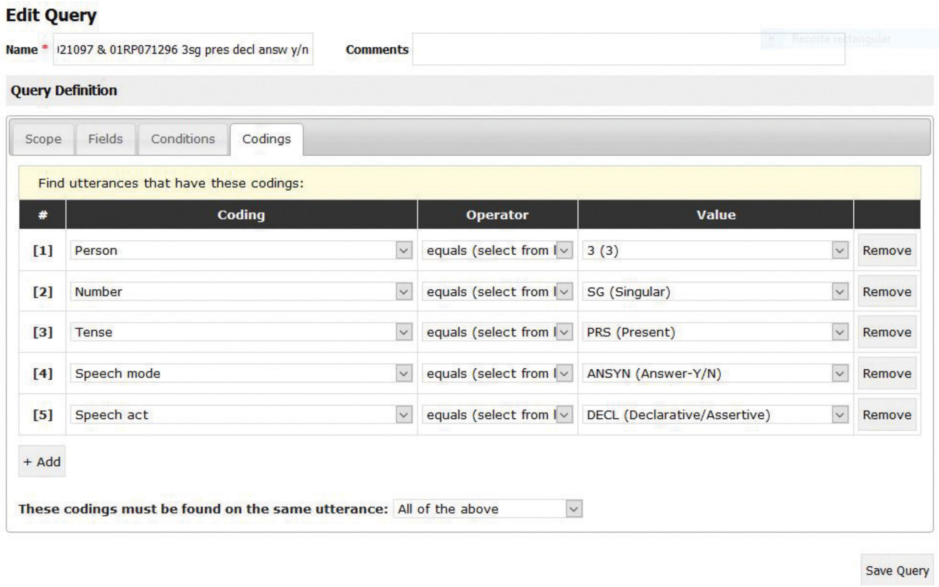
In the second query, we search for declarative sentences with third person singular present tense verbs produced as answers to *yes/no* questions. For this query we need to set coding titles and values to “speech act equals declarative,” “tense equals present,” “person equals 3,” “number equals singular,” and “speech form equals answer y/n.”

This query differs from the previous one on the items included in the codings tab, as shown in figure 9.8.

Our query produces the results shown in figure 9.9 for both the Spanish- and the English-speaking children.

It reports that the English-speaking subject, 04BG021097, produced one utterance out of 2,659 utterances and the Spanish-speaking subject, 01RP071296, produced two utterances out of 4,516 utterances that were declarative sentences with third person singular present tense verbs as answers to yes/no question during these sessions.[33]

Through such precise, targeted, context-specific analyses, all enabled by the DTA tool, a user can refine their research data, testing hypotheses by tailoring analyses to data with highly specific metadata characteristics across varying levels of linguistic analysis across languages, datasets, and projects. Once analyses such as these are standardized, large



Edit Query

Name * 21097 & 01RP071296 3sg pres decl answ y/n

Comments

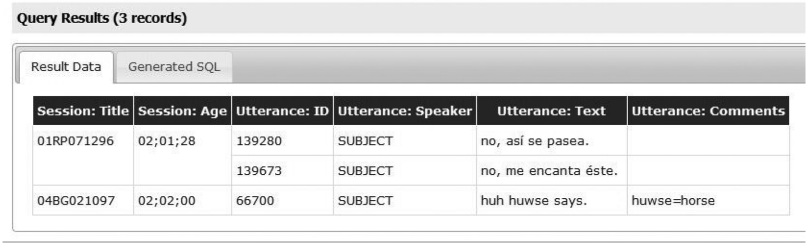
Query Definition

Find utterances that have these codings:

#	Coding	Operator	Value	
[1]	Person	equals (select from l	3 (3)	Remove
[2]	Number	equals (select from l	SG (Singular)	Remove
[3]	Tense	equals (select from l	PRS (Present)	Remove
[4]	Speech mode	equals (select from l	ANSYN (Answer-Y/N)	Remove
[5]	Speech act	equals (select from l	DECL (Declarative/Assertive)	Remove

These codings must be found on the same utterance: All of the above

Figure 9.8
Declarative sentences with 3rd person singular present tense verbs and answers to *yes/no* questions codings and values.



Session: Title	Session: Age	Utterance: ID	Utterance: Speaker	Utterance: Text	Utterance: Comments
01RP071296	02;01;28	139280	SUBJECT	no, así se pasea.	
		139673	SUBJECT	no, me encanta éste.	
04BG021097	02;02;00	66700	SUBJECT	huh huwse says.	huwse=horse

Figure 9.9
Declarative sentences with 3rd person singular present tense verbs and answers to *yes/no* questions results.

corpora can be studied and large populations can be subjected to systematically comparable analyses. Replication of our methodology on larger samples would likely produce evidence on the degree to which the results for this particular Spanish–English case comparison generalize children's language among larger groups, thereby dissociating individual differences within those groups. The DTA can be both flexible and effective for the sorting and display of language data, thus making the data relevant not only for more general developmental questions (e.g., development of utterance length in child language) but also for focused study of specific linguistic questions and hypothesis testing (e.g., in the case of omission of inflection in child language), including cross-subject and cross-linguistic comparison.

Moving to Linked Open Data

Once data are available through a digital infrastructure, they are available to scale to larger networks involving multiple users, and in addition to integrate with Linked Open Data frameworks. This data transformation would open each database to potential new analytic tools and to new suitable archiving means.

Unfortunately, at present in the language sciences even the various initiatives whose very intention is to cultivate research collaboration and wide dissemination through the use of interdisciplinary databases (e.g., the VCLA's DTA primary research tool and database, or the CHILDES database; see Bernstein Ratner and MacWhinney, this volume) are challenged to accomplish LOD interlinkage in their current form. Thus, import-export functions within and across systems remain limited today, as does scalability to a wider dissemination infrastructure (e.g., the university library; see Rieger, this volume).

One way to meet this challenge would be, first, to transform all existing data and resources into LOD and/or LOD-aware (i.e., Semantic Web, the vision of Berners-Lee 2006; see Pomerantz 2015, 153f for introduction) resources individually, and then to interlink them by means of suitable interconnecting references and mappings. As Chiarcos and Pareja-Lora and Moran and Chiarcos, both in this volume, show, there has been significant technical development of standards and ontologies to support the LOD transformation processes (see also Castano, Ferrara, and Montanelli 2006; Troncy et al. 2007; Trivellato et al. 2009; Moise and Netedu 2009; Pareja-Lora and Aguado de Cea 2010; Métral et al. 2010; Pareja-Lora 2012a, 2012b, 2013; Cavar et al. 2015). However, the development of the LOD cloud is still in its infancy, the conversion process remains quite complex, and many details still require in-depth discussion before solutions can be implemented.

Alternatively, researchers working with existing databases may begin by linking their systems to ontologies and standardization consistent with LOD. In information science "an ontology is a formal representation of the universe of things that exist in a specific domain" (see Pomerantz 2015 for an introduction). In each case, such linkage must work with formalization of the underlying conceptual structure of the system/database. The content of the existing database must be matched to that of the existing ontology, and the ontology must be standard conformant; in addition, in the case of linguistic LOD (LLOD), the ontology should be aware of and consistent with ISO/TC 37 standard categories, terminology, and/or knowledge, too (see Ide, and Warburton and Wright, both in this volume). Then, the formal categories in the established ontology must be related to the categories of the database (and they will be, accordingly, linked to standardized and/or standard-related categories).

In initial work, DTA tool resources (codings/annotations) have been linked to the standard conformant, ISO/TC 37 consistent OntoLingAnnot ontology (Pareja-Lora and Aguado de Cea 2010; Pareja-Lora 2012a, 2012b, 2013) for both DTA tool database metadata categories and data annotations (see Pareja-Lora, Blume, and Lust 2013). This process involves adaptation of both the formal ontological labels and those of the DTA tool, formalizing them as linguistic RDF triples (see Moran and Chiarcos, and Simons and Bird, both in this volume; see Pomerantz 2015, for introduction), as in the ontological model, in order to achieve standardization and transformation to LLOD. Subsequent conversion to the Semantic Web representation of the output of the ontology conversion lies ahead, however.

Such computational-linguistic mergers can benefit development of both areas. For example, once data are available through a digital infrastructure, they also become available for secondary analysis (e.g., for testing linguistic annotation systems created by natural language processing engineers working in the Semantic Web community [cf. Chiarcos, Nordhoff, and Hellmann 2012]). Making data more widely available increases the return that science as a whole garners from the significant investment of time as well as intellectual and material resources required for every piece of experimental or observational data that is collected, captured, and coded for analysis. In this community, researchers feel an increasingly urgent need for having research data and resources transformed into open, sharable, and interoperable data and resources by converting language data (that is, corpora, lexicons, and so on) into LLOD sets and/or graphs and language software resources (such as POS taggers or parsers) in LLOD-aware language resources. Linguistic Linked Data help formalize and make explicit common-sense knowledge in a way that satisfies the needs of the Web 3.0,[34] the Semantic Web, and/or the Web of Data. The merger process can lead to detection of inconsistencies and gaps both in existing ontologies and in existing databases such as those that result from the DTA tool. For the individual language researcher, or the community, such mergers can explicate standards that are necessary to support LOD processes. Several steps have already been taken toward this end, for instance within the European LIDER project,[35] and a number of best practices have already been identified for this purpose—for example, the LIDER project and the W3C's *Best Practices for Multilingual Linked Open Data Community Group*.[36] However, current research on the challenge posed by ontology conversion for the individual researcher needs to be supplemented with further work in order to ease and systematize this process, and possibly even to automatize it, if possible.

Testing Import–Export Functions for the DTA Tool

Development of LOD technology would support exchange of data across databases. Such exchange would empower research in general, merely by increasing the amount of data available to a single researcher and in addition increasing its comparability and generalizability. Also, more specifically, for example, an automatic import function would allow language transcripts in other databases to be imported to the DTA tool to allow situating their data in the metadata structure provided by the DTA tool, thus establishing its provenance and calibrating it for further use, and/or to exploit its advanced analytic functions, its collaborative structure, or its cross-linguistic and/or experimental data. An automatic export function of data in the DTA tool to other databases, such as CHILDES, for example, would integrate that data with the effective dissemination systems of CHILDES as well as with individual analyses that numerous researchers are conducting within that system.

Through the work of Cornell University's library technical consultant James Reidy, plus the support of the library's "tech innovation week" program, an initial exploration of a data import–export function to and from the DTA tool was conducted. An export function was first explored. The CHAT format (cf. endnote 37; also see Bernstein Ratner and MacWhinney, this volume) was selected, since it is a common format for transcribing child language data developed for the largest child language data resource, CHILDES (cf. endnote 5). The CLAN application (Computerized Language Analysis) is designed specifically to analyze data described in the CHAT format and has tools for checking the syntax of CHAT files.[37] In general, such import–export functions require the integration of the overall structure and labeling fields of each database. Figure 9.10 displays the names and relationships among the tables used by the DTA tool. The information used to construct the input for CLAN came from the subject, session, transcription, and utterance tables.

Two samples of natural speech data were selected (one Spanish, one English) as a target of exchange, since natural speech is the content of most data in the CHILDES data bank. Because CHAT is a text file format, the DTA tool's samples were written out as CHAT

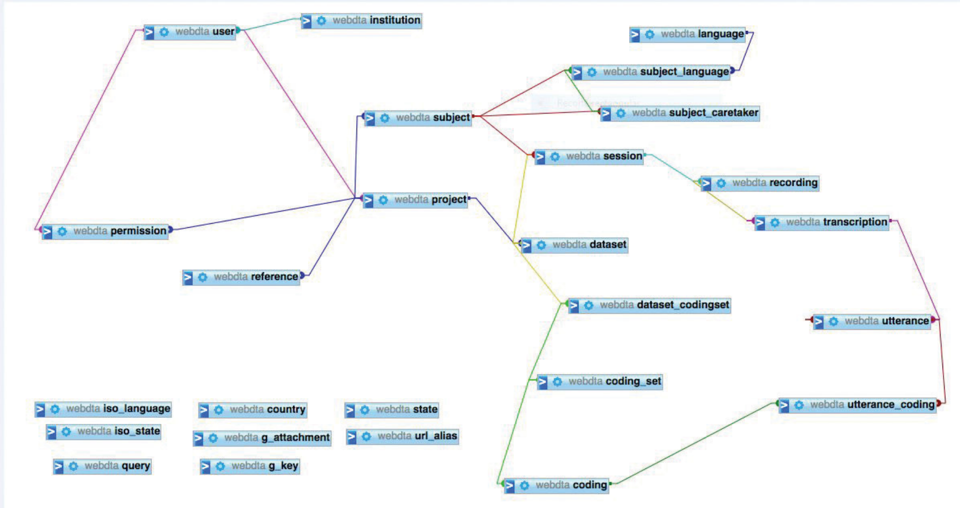


Figure 9.10
Names and relationship among tables used by the DTA tool.

files and then tested using CLAN's check command. While we will leave details on this LOD initiative to future reports, this initial process revealed several results that must guide future development. In general, the transfer is shown to be technically possible, yet the challenges that arise require human intervention going beyond current automaticity. The major challenges include the following:

1. If total import–export is sought, then the full array of fields will seek transfer; these will include both metadata and data fields. In the DTA tool, global codings (characterizing all projects) as well as individual dataset codings will be included. Since the extensive metadata fields surrounding the research data (layers above the transcript in the DTA tool, such as projects, datasets, sessions, subjects) are not replicated in other databases, including CHILDES, the first challenge is to calibrate metadata fields for transfer. The CHILDES metadata fields include some that the DTA tool does not (e.g., layout of child's home, religion, friends). At the same time, as introduced in Pareja Lora, Blume, and Lust 2013, because both CHILDES and the DTA tool have focused on child language data, they do share common or similar labels in the metadata codings they involve (although CHILDES lists metadata fields in its manual, while the DTA tool provides a structured interface for them in the database).
2. Another challenge involves a potentially complex mapping even of metadata fields alone. For example, the "creator" label corresponds to three labels in the DTA tool: "principal investigator," "additional investigators," and "assisting investigators." Some identical labels refer to different things across databases (Pareja Lora, Blume, and Lust 2013).

3. At the data level, issues of transfer include differences in encoding transcription across systems, which provide CHAT syntax errors. For example, CHAT encoding of the “Main Line” is quite different from the encoding used in the DTA tool’s utterance table’s text field.[38] CLAN wants one utterance per line and uses special characters at the end of an utterance. The DTA tool, by contrast, uses special characteristics within utterances and uses special markers to indicate properties of speech.
4. Display of CLAN error reports in the context of a transcript can identify for the researcher where human intervention is necessary for transfer, but this of course eliminates the automaticity sought.

One potential approach to the LOD framework would involve implementation of the DTA tool (and its established ontologies; Pareja-Lora, Blume, and Lust 2013) with Semantic Web technologies—a project that remains for the future. If the DTA data can be exported in a Semantic Web form and reimported into substantially the same structure as the original DTA data (that is, if lossless processes for the representation and/or transformation of these data can be implemented), then also data from other systems in the Semantic Web form (that is, providing LOD exports) could be imported into the DTA tool. Another option would be to develop analysis tools (like those already existing in the DTA tool, in CLAN, and in CHILDES) that can work directly on the Semantic Web form of the data. This would encourage further development of the ontology to cover the unique characteristics of each system and ultimately would provide an incentive to migrate existing data into the ontology.

Conclusions

Development of the DTA tool, which we have reviewed above, has pursued several of the promises offered to the interdisciplinary researcher in the language sciences in our increasingly digital and networked world. In turn, it has exposed challenges, which we address below.

Cybertools such as the DTA tool address the need for standardized yet flexible (and expandable) formats for data capture, including extensive metadata to establish data provenance. Such tools begin to address the challenge of interlinkage by providing the capacity to link data across projects and languages. Cybertools such as we have exemplified through the DTA tool also begin to address the challenges of data complexity. In the particular case of this tool, carefully sequenced screens for data and metadata entry and usable interfaces both guide and streamline the data-entry process. The design of this example tool has drawn on the insights of a community of scholars into the types of data and of metadata that are important to language acquisition research and to the relations among data points that can shed light on language acquisition questions, thereby documenting the critical importance of knowledge communities for cybertool development.

If extensively exploited, all the design properties of the DTA tool will begin to enable and empower collaborative research (Berners-Lee 2006, 2009; Wenger 1998; Wenger, McDermott, and Snyder 2002). The infrastructure of cybertools such as those exemplified by the DTA tool is designed to foster and enhance collaboration on the basis of shared materials, practices, and data, even at long distances, both within and across disciplines and languages. Such digitally enabled collaboration can, in turn, empower researchers working in the field of cognitive science to attack challenging new questions that require interdisciplinary approaches involving the language sciences.

At present, though, issues of scalability and interoperability must be confronted so that various efforts in the creation of language documentation and language acquisition repositories can be integrated, thus achieving further strength for each and for all, and finally realizing the vision of a Linked Open Data framework. As other chapters in this volume suggest, the technology for achieving such interoperability is advancing quickly.

Potential Expansions of the DTA Tool Application

Although our examples have drawn from first language acquisition, the DTA tool permits similar comparisons across many kinds of datasets—for example, data from first, second, and/or multilingual language acquisition, plus language delays, as well as language impairments in children or adults. Such comparisons within and across linguistic datasets from different populations and languages would be prohibitive without leveraging technology (see Blume et al. this volume).

A linguist can use the tool to enhance language documentation in general, including endangered languages (Lust et al. 2010; Bird 2011; Grenoble and Furbee 2010). The DTA tool would also be suitable for corpus linguistics, language pathology, language contact, and sociolinguistic studies, among others. In the field of speech and language pathology, a clinician could use the tool to track client progress among dimensions such as mean length of utterance (MLU), sentence type, or others through using the DTA tool. Recently, a study of language dissolution in an aging population evidencing prodromal Alzheimer's disease was built on an early study of the first language acquisition of relative clauses , which had been archived by and documented in the DTA tool to, first, replicate the experimental design (Flynn and Lust 1980) and methods with a new study of the elderly, and, second, conduct a critical comparison of results across children and both healthy and impaired elderly (e.g., Lust et al. 2015, 2017).

Acknowledgments

This project was supported by several funding sources: "Transforming the Primary Research Process through Cybertool Dissemination: An Implementation of a Virtual Center for the Study of Language Acquisition," National Science Foundation grant to María Blume and Barbara Lust, 2008, NSF OCI-0753415; "Planning Grant: A Virtual Center for Child Language Acquisition Research," National Science Foundation grant to Barbara Lust, 2003, NSF BCS-0126546; "Planning Information Infrastructure Through a New Library-Research Partnership," National Science Foundation Small Grant for Exploratory Research to Janet McCue and Barbara Lust, 2004–2006, #NSF 0437603; American Institute for Sri Lankan Studies, Cornell University Einaudi Center; Cornell University Faculty

Innovation in Teaching Awards, Cornell Institute for Social and Economic Research (CISER); New York State Hatch grant; and Grant Number T32 DC00038 from the National Institute on Deafness and Other Communication Disorders (NIDCD). Besides, the work of the second author has been partially supported by the projects RedR+Human (Dynamically Reconfigurable Educational Repositories in the Humanities, ref. TIN2014-52010-R) and CetrO+Spec (Creation, Exploration and Transformation of Educational Object Repositories in Specialized Domains, ref. TIN2017-88092-R), both financed by the Spanish Ministry of Economy and Competitiveness.

We thank James Gair for his continuous counsel and collaboration; application developers Ted Caldwell and Greg Kops (GORGES); consultants Cliff Crawford and Tommy Cusick; student research assistants Darlin Alberto, Gabriel Clandorf, Natalia Buitrago, Poornima Guna, Jennie Lin, Marina Kalashnikova, Martha Rayas Tanaka, Lizzeth Jensen, María Jiménez, and Mónica Martínez; The Cornell University Library's associate director, Oya Rieger, and its technical consultant, Jim Reidy; the Cornell Center for Advanced Computing; and finally, the many students at all participating institutions who helped us with comments and suggestions.

We gratefully acknowledge the collaboration of the other founding members of the Virtual Center for the Study of Language Acquisition: Marianella Casasola, Claire Cardie, and Qi Wang (Cornell University, USA); Elise Temple (The Nielsen Company, USA); Liliana Sánchez (Rutgers University at New Brunswick, USA); Jennifer Austin (Rutgers University at Newark, USA); YuChin Chien (California State University at San Bernardino, USA); and Usha Lakshmanan (Southern Illinois University at Carbondale, USA). We are thankful for the collaboration of scholars who are VCLA affiliates, including Sujin Yang (Ewha Womans University, South Korea); Gita Martohardjono, Valerie Shafer, and Isabelle Barrière (City University of New York, USA); Cristina Dye (Newcastle University, UK); Yarden Kedar (Beit Berl College, Israel); Joy Hirsch (Columbia University, USA); Sarah Callahan (Assessment Technology Incorporated, USA); Kwee Ock Lee (Kyungsung University, South Korea); R. Amritavalli (Central Institute of English and Foreign Languages, India); and A. Usha Rani (Osmania University, India).

Notes

1. Virtual Linguistic Lab (VLL). http://clal.cornell.edu/vll/.

2. Our proposals and the cybertool we introduce apply to first, second, and multiple language acquisition in child or adult. They have implications for the management and representation of language data in general.

3. Data Transcription and Analysis Tool (DTA). http://webdta.clal.cornell.edu.

4. https://www.w3.org/DesignIssues/LinkedData.html.

5. Child Language Data Exchange System (CHILDES). http://childes.talkbank.org/.

6. The Language Archive (TLA): https://tla.mpi.nl/tools/tla-tools/elan/.

7. Electronic Metastructure for Endangered Languages Data (E-MELD). http://emeld.org/index.cfm.

8. General Ontology for Linguistic Description (GOLD). http://linguistics-ontology.org.

9. Open Language Archives Community (OLAC). http://www.language-archives.org.

10. Open Linguistics Working Group (OWLG). http://linguistics.okfn.org.

11. This challenge is even more complex because the line between what constitutes "data" and what constitutes "metadata" is fluid (see Borgman 2015; Bender and Langendoen 2010; Pomerantz 2015).

12. Current efforts to leverage these capacities of technology are many; see Chiarcos, Nordhoff, and Hellmann (2012); see also DataStaR, an experimental data-staging repository that aims to enable collaboration and data sharing (Lowe 2009; Steinhart 2010; Khan et al. 2011; see also chapters in this volume).

13. An illustration is the observation that at young ages, children appear to omit auxiliary verbs in obligatory contexts, a phenomenon that calls into question the nature of children's early representations of language. For German child speech, Boser et al. (1991) proposed these apparent omissions as "phonetically null auxiliaries." Dye's (2011) analysis of child French using new, sensitive recording equipment provided relevant phonetic evidence in similar environments (see also Dye, C., Y. Kedar, and B. C. Lust, forthcoming).

14. Efforts to leverage technology in the standardization of data capture at the sentence level have been under way for some time (e.g., the glossing rules of Bickel, Comrie, and Haspelmath 2008).

15. Virtual Center for the Study of Language Acquisition (VCLA). http://vcla.clal.cornell.edu. Founding members are listed in the acknowledgments.

16. Foley's (1996) dissertation was written before the DTA existed in its current form. Her project has been included in the DTA to allow for comparisons with other projects. The data input is still in progress, so figure 9.1 reflects data from one age group (*n*=8) in the study.

17. See the site of supplemental materials for Blume and Lust (2017) at http://pubs.apa.org/books/supp/blume/?_ga=1.998898.2130472459.1479745044.

18. See the list of VLL founding institutions in the acknowledgments to this chapter.

19. The term *headless relatives* refers to the absence of the lexical head. There are ongoing debates on the valid representation of their syntactic structure; which are sometimes termed *free relatives*.

20. This *wh*-form, a syntactic "operator," is distinct in position from the complementizer *that*, which may also introduce English relative clauses (as in "the balloon *that* bumps Ernie"). For evidence supporting these structural analyses of lexically headed clauses and headless relative clauses in French and English, see the synthesis in Foley 1996.

21. Foley (1996) and Flynn et al., forthcoming, discuss the significance of these results.

22. Because these *global* codings can also be applied to natural speech data as well, they allow comparisons between natural speech and experimental data.

23. At present, the DTA tool can compute the following functions: average, minimum, maximum, sum, number of, standard deviation, and variance.

24. A fairly detailed set of English MLU criteria can be found in Blume and Lust (2017). The book's website contains supplemental materials, including the Spanish MLU criteria that Blume compiled after revising previous MLU criteria proposals available for Spanish at the time.

25. Figure 9.3 shows the video file as the main resource for the transcript. One can switch between resources and select to display audio or to download a PDF version of a previous transcript instead

(this is often used, for example, when one has only a handwritten version of the transcript created in the field). Researchers can select whether they want to have independent transcripts for each resource (audio, video, previous transcript) or to create a single transcript using all resources. The details of what resources were used for each transcript are specified on a previous screen. For the transcription conventions see Blume and Lust 2017, appendix A.

26. This notation means "years, months, days"—this child is two years and two months old.

27. Utterances that are complex sentences can be further analyzed and coded after they are divided in clauses in the Tagged Transcription screen.

28. Following the glossing rules of Bickel et al. (2008).

29. Coding sets can be reordered in the screen, so novice researchers can move particular sets to the bottom of the screen or keep them closed if desired. Most coding sets consist of drop-down lists or radio buttons, thus minimizing the possibility of typing errors.

30. To run an MLU query, the user selects the same scope as for the queries described below. Under "fields" the user selects the following:

Session: Title or Transcription: Title

Session: Age

Utterance: Speaker

Coding value = average

and selects "group by" next to Session: Title. The conditions are "Utterance: Speaker equals SUBJECT" and "Coding: Title equals Number of morphemes." Codings would be "Speech act does not equal Unclear" and "Number of morphemes does not equal 0."

31. The verb coding set marks whether the noninflected verb form would be allowed in adult grammars. In Blume (2002), it was concluded that to test issues of finiteness in child language, speech context must be evaluated at the same time as a unique utterance with an inflected or noninflected verb. On the bases of the context, certain noninflected verb forms were identified as non-adult-like in both Spanish and English.

32. These IDs specify the session number, child initials, and birth date (see Blume and Lust 2017 and Blume and Lust 2012a for discussion of subject IDs).

33. One can conduct a similar query adding "coding title" and "coding value" to the fields to see all the codings applied to the utterances shown in the query's results.

34. http://en.wikipedia.org/wiki/Web_2.0#Web_3.0.

35. LIDER Project. http://lider-project.eu/lider-project.eu/index.html.

36. https://www.w3.org/community/bpmlod/.

37. https://talkbank.org/manuals/CHAT.pdf.

38. https://talkbank.org/manuals/CHAT.pdf, p. 42.

References

Abney, S. 2011. "Data-Intensive Experimental Linguistics." *Linguistic Issues in Language Technology* 6. http://elanguage.net/journals/lilt/article/view/2578.

Atkins, D. E., K. K. Droegemeier, S. I. Feldman, H. García-Molina, M. L. Klein, D. G. Messerschimdtt, P. Messina, et al. 2003. "Revolutionizing Science and Engineering: Report of the National

Science Foundation Blue-Ribbon Advisory Panel on CyberInfrastructure," January 2003. http://www.nsf.gov/cise/sci/reports/atkins.pdf.

Bender, E., and D. T. Langendoen. 2010. "Computational Linguistics in Support of Linguistic Theory." *Linguistic Issues in Language Technology*. CSLI Publications.

Berman F., and H. E. Brady. 2005. "Workshop on Cyberinfrastructure for the Social and Behavioral Sciences Final Report." http://ucdata.berkeley.edu/pubs/CyberInfrastructure_FINAL.pdf.

Berners-Lee, T. 2006. "Linked Data." http://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, T. 2009. "The Next Web." http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html.

Bickel, B., B. Comrie, and M. Haspelmath. 2008. "Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-Morpheme Glosses." Accessed April 15, 2017. http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Bird, S. 2011. "Bootstrapping the Language Archive: New Prospects for Natural Language Processing in Preserving Linguistic Heritage." *Linguistic Issues in Language Technology* 6. http://elanguage.net/journals/lilt/article/view/2580.

Bird, S., and G. F. Simons. 2003. "Seven Dimensions of Portability for Language Documentation and Description." *Language* 79:557–582.

Blume, M. 2002. "Discourse-Morphosyntax Interface in Spanish Non-Finite Verbs: A Comparison between Adult and Child Grammars." Ph.D. diss., Cornell University.

Blume, M., S. Flynn, and B. C. Lust. 2012. "Creating Linked Data for the Interdisciplinary International Collaborative Study of Language Acquisition and Use: Achievements and Challenges of a New Virtual Linguistics Lab." In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata,* edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 85–96. New York: Springer.

Blume, M., C. Foley, J. Whitlock, S. Flynn, and B. C. Lust. 2014. "A New Data Management Cybertool Supports Cross-Linguistic Collaborative Research and Student Training." In *Boston University Conference on Language Development*, edited by W. Orman and M. J. Valleau. Somerville, MA: Cascadilla Press.

Blume, M., and B. C. Lust. 2012a. "Data Transcription and Analysis (DTA) Tool User's Manual" (with the collaboration of Shamitha Somashekar and Tina Ogden). http://webdta.clal.cornell.edu.

Blume M., and B. C. Lust. 2012b. "First Steps in Transforming the Primary Research Process Through a Virtual Linguistic Lab for the Study of Language Acquisition and Use: Challenges and Accomplishments." *Journal of Computational Science Education* 3:34–46.

Blume, M., and B. C. Lust. 2017. *Research Methods in Language Acquisition: Principles, Procedures and Practices*. Mouton de Gruyter and American Psychological Association.

Borgman, C. L. 2007 *Scholarship in the Digital Age*. Cambridge, MA: MIT Press.

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Boser, K., B. Lust, L. Santelmann, and J. Whitman. 1992. "The Syntax of CP and V-2 in Early German Child Grammar: the Strong Continuity Hypothesis." Proceedings of the North Eastern Linguistics Association 22:51–66. Amherst: University of Massachusetts.

Castano, S., A. Ferrara, and S. Montanelli. 2006. "Ontology-based Interoperability Services for Semantic Collaboration in Open Networked Systems." In *Interoperability of Enterprise Software and Applications*, edited by D. Konstantas, J. P. Iourrières, M. Léonard, and N. Ioudjlida, 135–146. Berlin: Springer.

Cavar, D., M. Cavar, and D. T. Langendoen. 2015. "Gold." Paper presented at the Workshop on the Development of Linguistic Linked Open Data (LLOD) Resources for Collaborative Data Intensive Research in the Language Sciences, at the Linguistic Society of America Summer Institute, University of Chicago, July 26, 2015.

Chiarcos, C., S. Nordhoff, and S. Hellmann, eds. 2012. *Linked Data in Linguistics. Representing Language Data and Language Metadata*. Berlin: Springer.

CHILDES. http://childes.talkbank.org/.

DataStaR. https://sites.google.com/site/datastarsite/.

Data Transcription and Analysis Tool. http://webdta.clal.cornell.edu.

Dye, C. D. 2011. "Reduced Auxiliaries in Early Child Language: Converging Observational and Experimental Evidence from French." *Journal of Linguistics* 47:301–339.

Dye, C. D., C. Foley, M. Blume, and B. C. Lust. 2004. "Mismatches between Morphology and Syntax in First Language Acquisition Suggest a 'Syntax-first' Model." In BUCLD 28 Online Proceedings Supplement, edited by A. Brugos, L. Micciulla, and C. E. Smith. http://www.bu.edu/bucld/proceedings/supplement/vol28/.

Dye, C., Y. Kedar, and B. C. Lust. Forthcoming. "From Lexical to Functional Categories: New Foundations for the Study of Language Development." In The Role of Grammatical Words in Syntactic Development, edited by Anat Ninio. *First Language*, Special issue.

E-MELD: Electronic Metastructure for Endangered Languages Data. (forthcoming). http://emeld.org/index.cfm.

Farrar, S. O. and D. T. Langendoen. 2003. "A Linguistic Ontology for the Semantic Web." *GLOT International* 7:97–100.

Flynn, S., and C. Foley. 2004. "On the Developmental Primacy of Free Relatives." *MIT Working Papers in Linguistics* 48:59–69.

Flynn, S., C. Foley, J. Gair, and B. Lust. 2005. "Developmental Primacy of Free Relatives in First, Second and Third Language Acquisition: Implications for Their Syntax and Semantics." Paper presented at Linguistic Association of Great Britain Annual Meeting 2005, University of Cambridge, September 2, 2005.

Flynn, S., C. Henderson, C. Foley, and B. C. Lust. Forthcoming. "On the Acquisition of Headedness in Relative Clauses: Syntax Is Independent of Semantics."

Flynn, S., and B. C. Lust. 1980. "Acquisition of Relative Clauses: Developmental Changes in Their Heads." In *Cornell Working Papers in Linguistics* 1:33–45. Ithaca, NY: Department of Modern Languages and Linguistics, Cornell University.

Foley, C. 1996. "Knowledge of the Syntax of Operators in the Initial State: The Acquisition of Relative Clauses in French and English." Ph.D. diss., Cornell University.

GOLD (General Ontology for Linguistic Description). http://linguistics-ontology.org.

Grenoble, L. A., and N. L. Furbee, eds. 2010. *Language Documentation: Practice and Values*. Amsterdam: John Benjamins.

Khan, H., B. Caruso, J. Corson-Rikert, D. Dietrich, B. Lowe, and G. Steinhart. 2011. "DataStaR: Using the Semantic Web Approach for Data Curation." *International Journal of Digital Curation* 2:209–221.

King, G. 2011. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331:719–721.

King, T. H. 2011. "(Xx*-) Linguistics: Because We Love Language." *Linguistic Issues in Language Technology* 6. http://elanguage.net/journals/lilt/article/view/2585.

Langendoen, D. T., B. Fitzsimmons, and E. Kidder. 2005. "The GOLD Effort So Far." Paper presented at the 2005 E-MELD Workshop on Ontologies in Linguistic Annotation, Harvard University, July 1, 2005.

Language Archives, The. https://tla.mpi.nl/tools/tla-tools/elan/.

Lave, J., and E. Wenger, 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Lowe, B. 2009. "DataStaR: Bridging XML and OWL in Science Metadata Management." *Metadata and Semantics Research* 46:141–150.

Lust, B. C., S. Flynn, M. Blume, E. Westbrooks, and T. Tobin. 2010. "Constructing Adequate Language Documentation for Multi-faceted Cross Linguistic Data: A Case Study from a Virtual Center for Study of Language Acquisition." In *Language Documentation: Theory, Practice, and Values*, edited by L. Grenoble and N. L. Furbee, 89–108. Amsterdam: John Benjamins.

Lust, B. C., S. Flynn, and C. Foley. 1996. "What Children Know about What They Say: Elicited Imitation as a Research Method." In *Methods for Assessing Children's Syntax*, edited by D. McDaniel, C. McKee, and H. Smith Cairns, 55–76. Cambridge, MA: MIT Press.

Lust, B. C., S. Flynn, J. C. Sherman, J. Gair, J. Whitlock, C. Cordella, C. Henderson, et al. 2015. "Reversing Ribot: Does Regression Hold in Normal Aging or Prodromal Alzheimer's Disease?" *Brain and Language* 143:1–10.

Lust, B., S. Flynn, J. Cohen Sherman, C. Henderson, J. Gair, M. Harrison, and L. Shabo. 2017. "On the Biological Foundations of Language: Recent Advances in Language Acquisition, Language Deterioration and Neuroscience Begin to Converge." *Biolinguistics: Special Issue Celebrating Biological Foundations of Language* 11:115–137.

Lust, B. C., C. Foley, and C. D. Dye. 2015. "The First Language Acquisition of Complex Sentences." In *The Cambridge Handbook of Child Language,* 2d ed., edited by E. Bavin and L. Naigles, 298–323. New York: Cambridge University Press.

Lust, B. C., B. Lowe, S. Flynn, M. Blume, J. Corson-Rikert, and J. McCue. 2005. "Searching Interoperability between Linguistic Coding and Ontologies for Language Description: Language Acquisition Data." http://www.emeld.org/workshop/2005/proceeding.html.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk, 3d ed*. Mahwah, NJ: Lawrence Erlbaum Associates.

Métral, C., R. Billen, A. F. Cutting-Decelle, and M. van Ruymbeke. 2010. "Ontology-based Approaches for Improving the Interoperability between 3d Urban Models." *Journal of Information Technology in Construction* 15:169–182.

Moise, G., and L. Netedu. 2009. "Ontologies for Interoperability in the eLearning Systems." Universităţii Petrol–Gaze din Ploieşti, vol. LXI: 75–88.

NSF (National Science Foundation). 2007. "Cyberinfrastructure Vision for 21st Century Discovery." Accessed April 15, 2017. http://www.nsf.gov/pubs/2007/nsf0728/.

OLAC (Open Language Archives Community). http://www.language-archives.org.

OWLG (Open Linguistics Working Group). http://linguistics.okfn.org.

Pareja-Lora, A. 2012a. "OntoLingAnnot's Ontologies: Facilitating Interoperable Linguistic Annotations (Up to the Pragmatic Level)." In *Linked Data in Linguistics: Representing Language Data and Metadata*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 117–127. Heidelberg: Springer.

Pareja-Lora, A. 2012b. “OntoLingAnnot’s LRO: An Ontology of Linguistic Relations.” In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012).* Madrid, June 2012, 49–64. http://www.oeg-upm.net/tke2012/proceedings, paper 04.

Pareja-Lora, A. 2013. “The Pragmatic Level of OntoLingAnnot’s Ontologies and Their Use in Pragmatic Annotation for Language Teaching.” In *Languages for Special Purposes in the Digital Era*, edited by J. Arús, M. E. Bárcena, and T. Read, 323–344. Springer.

Pareja-Lora, A., and G. Aguado de Cea, 2010. “Modeling Discourse-Related Terminology in OntoLingAnnot’s Ontologies.” In *Proceedings of the TKE 2010 Workshop Establishing and Using Ontologies as a Basis for Terminological and Knowledge Engineering Resources*.” August 2010, Dublin.

Pareja-Lora, A., M. Blume, and B. Lust. 2013. “Transforming the Data Transcription and Analysis Tool Metadata and Labels into a Linguistic Linked Open Data Cloud Resource.” In Linked Data in Linguistics second workshop, edited by P. Cimiano, J. McCrae, C. Chiarcos, and T. Declerck, Pisa, Italy, September 23, 2013. *https://www.aclweb.org/anthology/W/W13/W13-55.pdf.*

Phillips, C. 1995. “Syntax at Age Two: Cross-Linguistic Differences.” *MIT Working Papers in Linguistics* 26: 352–382.

Pomerantz, J. 2015. *Metadata*. MIT Press Essential Knowledge Series, Cambridge, MA: MIT Press.

Radford, A. 2004. *English Syntax: An introduction.* Cambridge: Cambridge University Press.

Simons, G. F., S. O. Farrar, B. Fitzsimons, W. D. Lewis, D. T. Langendoen, and H. Gonzalez. 2004. “The Semantics of Markup: Mapping Legacy Markup Schemas to a Common Semantics.” In *Proceedings of the 4th Workshop on NLP and XML (NLP XML-2004): Held in Cooperation with ACL-04*, edited by Nancy Ide, Graham Wilcock, and Laurent Romary, 25–32. Stroudsburg, PA: Association for Computational Linguistics.

Somashekar, S. 1999. “Developmental Trends in the Acquisition of Relative Clauses: Cross-Linguistic Experimental Study of Tulu.” Ph.D. diss., Cornell University.

Steinhart, G. 2010. “DataStaR: A Data Staging Repository to Support the Sharing and Publication of Research Data.” 31st Annual IATUL Conference: The Evolving World of e-Science: Impact and Implications for Science and Technology Libraries, West Lafayette, Indiana, June 2010. http://docs.lib.purdue.edu/iatul2010/conf/day2/8/.

Trivellato, D., F. Spiessens, N. Zannone, and S. Etalle, S. 2009. “Reputation-Based Ontology Alignment for Autonomy and Interoperability in Distributed Access Control.” In *International Conference on Computational Science and Engineering 2009,* 3:252–258. IEEE.

Troncy, R., Ó Celma, S. Little, R. García, and C. Tsinaraki. 2007. “Mpeg-7 Based Multimedia Ontologies: Interoperability Support or Interoperability Issue.” In *1st International Workshop on Multimedia Annotation and Retrieval Enabled by Shared Ontologies*, 2–15.

U.S. Office of Science and Technology Policy. 2013. February 22 memo. http://www.whitehouse.gov/administration/eop/ostp/library/publicaccesspolicy.

VCLA (Virtual Center for the Study of Language Acquisition), http://vcla.clal.cornell.edu.

VLL (Virtual Linguistic Lab), http://clal.cornell.edu/vll/.

Wenger, E. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.

Wenger, E., R. McDermott, and W. Snyder. 2002. *Cultivating Communities of Practice: A Guide to Managing Knowledge*. Cambridge: Harvard Business School Press.

Zagona, K. 2001. *The Syntax of Spanish*. Cambridge: Cambridge University Press.

10 Challenges for the Development of Linked Open Data for Research in Multilingualism

María Blume, Isabelle Barrière, Cristina Dye, and Carissa Kang

Introduction

The study of multilinguals is fundamental for linguistic research, since multilinguals constitute the majority of the world population as well as a growing proportion of the population in many countries (McCabe et al. 2013; Gambino, Acosta, and Grieco 2014; Special Eurobarometer 386, 2012). We use the term *multilingual* to refer to speakers who know more than one language to a variable extent, regardless of when they learned those languages (thus encompassing simultaneous and sequential bilinguals, as well as second-language speakers/learners and heritage speakers).

The multilingual brain is dealing with more than one linguistic system, and thus theories of language structure and cognitive models of language development or processing must account for language use, processing, and acquisition by all people who know more than one language. The language abilities of multilinguals change throughout their lifetime, so our data need to capture differences in a person's ability through time, including language attrition when or if it occurs. Studies on bilingualism, multilingualism, second-language acquisition, and language attrition have grown exponentially in the last decades, and their data need to be accessible and comparable so that all the research community can benefit from it.

With these facts in mind, we discuss three major issues related to research with multilingual populations:

- Requirements for conducting research with multilingual populations
- Challenges for the development of Linguistic Linked Open Data (LLOD) in the field of multilingual acquisition
- Capacities and needs of any primary research tool that would allow us to achieve the vision of LLOD

Requirements for Conducting Research with Multilingual Populations

Several methodological issues arise in doing research with human participants in the field of linguistics (Blume and Lust 2017). However, working with multilingual participants creates additional challenges.

Complexity Inherent in Multilingual Population

Most people become multilingual because their circumstances force them to do so. These different circumstances can be summarized as follows[1] (Austin, Blume, and Sánchez 2015, 39):

- Individuals who are multilingual from birth either speaking two languages at home or one at home and one outside the home
- Early multilinguals who start learning a second language sometime after birth but still during childhood, typically speaking one language at home and one outside the home
- Individuals who learned a second language in adulthood and speak it mostly for work-related activities
- Individuals who, as a result of immigration, must learn a second language to survive in the new country, or who spoke a minority language in their own country but must learn the dominant language in their new country

Even within groups, multilingual speakers differ greatly in many respects. They may range from monolingual speakers having limited exposure to a second language (e.g., a few hours per week in a classroom), to more fully multilingual speakers (e.g., people who learned both languages simultaneously in childhood, using both frequently in everyday life across various situations). A speaker's proficiency may also change across different contexts (Fishman 1965) and throughout the speaker's lifetime (more so than that of a monolingual speaker), requiring assessments across various situations and at multiple points in their language development.

Determining the nature of a participant's multilingualism is fundamental for research since it has effects on such important areas as further linguistic development, cognition, and literacy.

Challenges for Research Posed by Population Complexity

Because many complex factors account for a multilingual person's language profile, it is often challenging to select individuals for study who have only some specific characteristics that a researcher wishes to compare, or to form groups of speakers of similar characteristics that one can then compare to different groups (in the same study or across studies). Detailed metadata must be carefully collected and documented to allow for such comparisons.

Those who are even considered to be possible participants change across studies. Depending on the type of research and how researchers define multilingualism, types of participants who are recruited may differ considerably. Some studies may call a participant *multilingual*, for example, only if he or she is a simultaneous multilingual—someone who learned two or more languages from birth with only a few days of difference between the beginning of exposure to each language (De Houwer 2009). Others will count students in their first stages of classroom-only exposure to a second language as being *multilingual*.

This may be largely attributed to the fact that there is no single definition in the field nor any clear set of criteria for deciding who is a multilingual speaker (Hamers and Blanc 1989; Grosjean 2010; Mackey 2012), arising from the complexity of the multilingualism phenomenon. Criteria used to characterize multilingual speakers include psychological ones (such as the degree of competence in one language versus another; Lambert 1955); the domains of competence (spoken and/or written production, oral comprehension and/or reading abilities; Bialystok 2007); and sociological ones, such as the contexts of use of a language and whether it was acquired in a naturalistic context or a formal setting (Fishman 1965). To complicate the matters further, terms commonly used to classify speakers in the literature, such as *balanced*, *dominant, native,* or *beginner*, refer to different concepts and are related to the different types of criteria, making comparison across studies less direct (Flege, MacKay, and Piske 2002; Genesee 1989; Genesee, Nicoladis, and Paradis 1995; Hamers and Blanc 1989).

Although some criteria undoubtedly exist in our field, not all relevant factors are systematically taken into account (for example, speakers are classified according to age of acquisition, but patterns of use may not be considered). At other times, speakers are carefully selected, though the criteria for selection are not completely or clearly detailed in publications (Grosjean 2008, 2011; Thomas 1994). It may not be realistic to expect all researchers to agree on the exact definitions of terms or to have them list in their research articles every last criterion used for classifying speakers, largely because of space limitations. However, the value of each study data can be incremented if researchers make this detailed information available online, so that other researchers can decide whether the population studied fits the profile they are looking for, either for further research with the same data or for comparison with other data.

To be able to compare groups of speakers, researchers need to control several potentially confounding factors in order to conclude that a speaker's multilingualism modulates, for example, the use of a particular linguistic structure or leads to a proposed cognitive difference. Two such factors are the context of acquisition and the type or level of multilingualism involved.

Context of Acquisition

To be able to establish the context of acquisition of an individual's languages, a researcher needs to have information on the speaker's language history—such as "Which languages has the participant acquired?" or "When and how were the languages acquired?" Age of acquisition is a good predictor of further language proficiency, with people who acquire a second language early usually outperforming speakers who acquired the language later in terms of linguistic abilities. The status of the language in the speaker's society is also important. Speakers tend to use and maintain languages that the majority of the population speaks and that their societies consider important more than they use and maintain minority languages, often because of a lack of educational resources and opportunities to use the language in daily life. The relationship between a speaker's languages (e.g., How closely related are they? Which aspects of the language systems are similar and which are not?) should be also taken into account. A language that is closely related to a person's native language may be easier to acquire for a second language speaker than a more distantly related one (Grosjean 2008, 2011).

This information, and a participant's biographical data (such as sex, age, and socioeconomic status), make it possible to begin to assemble a language profile for the participant.

Type/Level of Acquisition

Information is also needed on the speaker's knowledge and use of each of his or her languages. This information is relevant to research for the reasons listed here, among others:

- *Language proficiency in the four skills (speaking, comprehension, reading, and writing) in each language:* Speakers may be similar in their comprehension skills but quite different in their expressive skills; some highly competent speakers may even be illiterate, and literacy has been shown to affect language processing.
- *Function of languages:* Which languages are used for what purposes? In what context and to what extent is each language used? Some speakers may have an extremely developed home-related lexicon in a language but not an academic one, or they may be able to have conversations about certain topics but not others. This may affect their performance on certain linguistic tests or their self-perception as multilingual speakers.
- *Language stability:* Are one or several languages still being acquired? Has a certain level of language stability been reached? In the past, wrong conclusions on the cognitive or linguistic capacities of multilingual speakers have often been reached when not taking into account that the subjects were incipient language learners of the language used for testing them.
- *Language modes:* This refers to the duration and frequency spent by the participant in both monolingual and multilingual modes. The mode may affect performance, especially in processing tasks. A speaker with less code-switching experience (i.e., alternating between more than one language) may provide very different answers to a study searching for syntactic or pragmatic constraints on code-switching than would a more experienced one.

Most studies gather information on language proficiency. However, language proficiency is not always understood or operationalized in the same way, and different instruments are frequently used to measure it. For example, some studies measure language competence (i.e., knowledge of the grammatical rules of a language), while others assess proficiency or communicative competence (i.e., the knowledge and ability to use language in socially acceptable ways, including grammatical, sociolinguistic, strategic, and discourse competence; Canale and Swain 1980; Canale 1983). Researchers are now more aware of this difference, and studies today tend to be more precise on their definition of competence.

To enable comparisons across multilingual speakers and groups, researchers need data on how their level of multilingualism was determined, including whether competence or proficiency were studied, the specific measures and tests used in assessing them, the task modality (e.g., comprehension or production), and the linguistic domain tested (e.g., vocabulary, grammar, pronunciation).

Sometimes speakers' proficiency is never directly measured—for instance, studies with L2 (second-language) learners are frequently conducted in formal settings (universities and schools) and course level is often used as a proxy measure for the speaker's proficiency (Thomas 1994). The problem with this approach is that courses that are officially at the same level (say, intermediate) may not actually be equally demanding at different institutions or across languages in the same institution.

When studies do gather independent data, questionnaires are frequently used. The questionnaires vary across labs in terms of length and type of information asked. Some are very short (approximately 10 questions), while others are much longer.[2] Although shorter questionnaires may be more practical, it is sometimes challenging to tell whether the results of a given study will generalize to other groups of multilingual speakers without detailed information.[3] Moreover, not all questionnaires of similar length ask the exact same questions about the speaker's acquisition, proficiency, and use.

Parental questionnaires have long been used as a measure of child language development (Gutierrez-Clellen and Kreiter 2003; Squires, Bricker, and Potter 1997; Thordaardottir and Weismer 1996), a recent study found that a more precise estimate of grammar can be achieved by adding a direct observation measure to the child's evaluation. In the study by Lust et al. (2014), two Korean-dominant children who were four years of age with Korean as their L1 (first language) and English as their L2 were assessed through a questionnaire and also an elicited imitation task. The parental reports and general linguistic histories predicted similar proficiency for the two children. However, in the experimental task, one child demonstrated a more developed level of grammar in his production in both of his languages than the other one. Thus, children who seem to be very similar according to parental reports can differ tremendously on their performance in experimental tasks both in the L2 and in the L1.[4]

While studies sometimes use standardized instruments to assess the development of linguistic abilities of the speaker, most such instruments exist strictly in English or only in a few well-studied languages (although a collection of instruments for research on second language acquisition can now be found through IRIS).[5] New instruments (or translations of existing instruments) that are reliable have proven difficult to create (e.g., Esquinca, Yaden, and Rueda 2005; Gathercole 2010; Paradis, Emmerzael, and Duncan 2010; Peña 2007), and it can take years to validate and norm them (Alcock et al. 2015). Some instruments measure only some aspects of linguistic knowledge—for example, vocabulary (e.g., Peabody Picture Vocabulary Test, Dunn and Dunn 2007). Most are normed on the basis of monolingual speakers (Espinosa and García 2012; Barrière 2014) and, as is well-known, bilinguals are not two monolinguals in one person and therefore cannot be compared directly to monolinguals (Grosjean 1989; Barac et al. 2014; and Sánchez 2015, among others).

Awareness of these differences among speakers and acknowledgment of the importance of having detailed information on their evaluation or classification has grown with the development of the field. This awareness is leading researchers to use more than one method to select and evaluate participants, as well as to collect more-careful metadata on each one. This is good for the field but it increments the amount of data and metadata that we professionals need to collect, store, and share.

Additional challenges may arise when researchers work with less-studied languages, in multilingual areas. It may often be the case that at least one of these less-studied languages is acquired by children and used in contexts where they constitute a minority language (Baker, van den Bogaerde, and Woll 2008). For instance, in New York City, 50% of children use a language other than English at home, including Haitian Creole, Yiddish, African Languages, Tagalog, Urdu, or Gujarati, and many others (García, Zakharia, and Otcu 2013, 13).

Even when both languages are well documented in adults, they may be less so for children. For example, while the use of both Spanish and English by Spanish-speaking adults in New York City has been documented (e.g., Otheguy and Zentella 2012 and references therein), little is known about the contextual factors that affect the acquisition of both languages in multilingual children. Barrière et al. (2015) investigated the acquisition of subject–verb agreement markers in English and Spanish by low socioeconomic status (SES) children of Mexican descent with Spanish as an L1: Their speakers were homogeneous with respect to the variety of Spanish they were acquiring, ensuring that the effects of bilingual acquisition were not confounded with dialectal variation in Spanish that impacts the speed and pattern of acquisition of Spanish grammatical inflections (e.g., Miller and Schmitt 2010). It was, however, difficult to determine the characteristics of the variety of English (such as Mainstream American English versus Chicano English or African American English or other Caribbean English) spoken by each participant. That determination was needed because different language varieties exhibit different norms regarding the third-person singular marker, and also because monolingual English-speaking children enrolled in the same preschools as their bilingual or trilingual colleagues perform differently on experimental tasks. That difference arises depending on the variety of English they are acquiring: Only preschoolers who are acquiring Mainstream American English (but not those acquiring other varieties, such as African American English or Jamaican

English) show evidence of comprehension in a video matching task that requires the exclusive use of the third-person singular–*s* to determine number of participants (examples of stimuli: *the boy skip**s*** versus *the boy**s** skip**ø***; Barrière et al. 2016).

The challenge of determining participants' language variety is significantly exacerbated when the languages to which the children are exposed to have not been well documented. This is the case of the Hasidic Yiddish-speaking community—a rapidly increasing population in two areas of Brooklyn—whose members speak varieties that come from three distinct areas in Eastern Europe that are now in contact both with one another and with English (Barrière 2010).

Studies conducted on multilinguals also frequently gather information on the attitudes that such speakers and their communities have about the languages they speak, attitudes that are relevant for explaining language dominance. Language preference has been shown to contribute to children's developing language abilities (Armon-Lotem et al. 2014). Some studies require more specific information; for example, Kang, Martohardjono, and Lust (unpublished manuscript) asked participants to self-rate the frequency of their daily language-mixing, the extent of their multilingualism, and even their attitudes toward code-switching, so as to investigate how code-switching attitudes and habits relate to code-switching fluency. Although language preference and code-switching behavior may affect multilingual development, they are rarely included in participant profiles.

All the previous examples illustrate how multilingual research requires extensive and detailed metadata to be gathered from each participant, which then need to be made accessible and searchable. The main issue is that more variability occurs among multilingual speakers' proficiencies than among those of monolingual speakers, and therefore researchers need to be able to describe multilingual participants in precise ways that are both meaningful and consistent across the field. These extensive data then must be documented and shared so that they benefit the wider research community.

Development of Linguistic Linked Open Data (LLOD)

Metadata

All the aspects of conducting research with multilingual populations discussed in the section "Requirements for Conducting Research with Multilingual Populations" point to the necessity of gathering extensive metadata on each participant before even testing them on the particular linguistic aspect of interest—metadata that are more extensive than for monolinguals. These metadata need to include not only the biographical and language context data mentioned above but also the specific measures used to classify the speakers' language abilities. Furthermore, multiple measures may be associated with each participant, since his or her abilities may change with age or development.

Most important, all these metadata must be linked to the particular data of the participant being studied.

Advantages of Accessible Extensive Metadata

Published studies should provide as much information as possible about their participants, the criteria used to classify multilinguals in various groups, and the language assessment tools used in the study; however, this is not always possible owing to length constraints. Having this information available online, then, would greatly facilitate research and calibration across studies.

Since it is often difficult to identify participants with a shared profile, studies of multilingual populations usually have small sample sizes. A tool that allows researchers to conduct meta-analysis studies (e.g., combining data collected from studies that employed a given task, or studies that focused on the development of a particular grammatical element) would certainly be advantageous, yet such analyses can only be properly conducted if we have access to exhaustive metadata for all studies.

Challenges

This extensive metadata documentation is now partially possible through some online tools (e.g., the DTA tool,[6] the Language Archive,[7] the Open Science Framework [OSF][8]), although the metadata, while available, are not always searchable automatically for less technically proficient researchers and the tools used to create them are often incompatible.

Gathering such detailed and often-personal data has the advantage of allowing us researchers to build an accurate linguistic profile of a multilingual speaker, but this brings with it the challenge of protecting the individual's identity, especially since multilingual speakers may come from minority and at-risk populations.

Data Challenges

In many cases, metadata and primary data either are not online or are not searchable; for example, the Electronic World Atlas of Varieties of English (EWAVE),[9] classifies varieties of English according to whether it is an L1 or L2 for the speakers, yet it provides no metadata on the informants. Many studies of multilingualism, for example, gather data and metadata through questionnaires. Although the results of a given study may be available online, the questionnaires themselves often are not, and at best they are attached as PDF forms to participants' metadata. This situation creates difficulties for comparison, calibration, and replication of studies.

Data Markup Challenges

Cross-Linguistic Differences

The main problem that multilingual data present is precisely that of being *multilingual*. Structures require an additional level of coding, indicating which language they belong to (in those cases where the researcher can even confidently decide the language). While this may be easy to do for independent words or one-language utterances, it can be more challenging in multi-language utterances and in utterances where words themselves contain morphemes from more than one language.

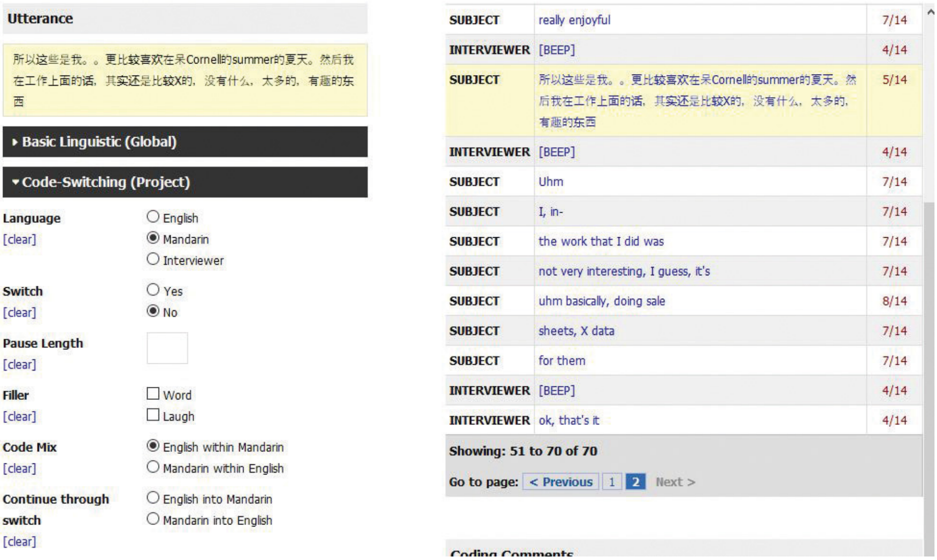
Enabling the cross-linguistic analyses needed to compare a multilingual speaker's two or more languages requires a rich markup capacity. Coding systems for the two or more languages need to be available for the researcher, and specific coding conventions may also need to be created, depending on the languages involved, since some phenomena common in the speech of a number of multilingual communities may be rare or nonexistent in others. For example, when analyzing the imitation of relative clauses in three languages (Flynn and Lust 1981; Foley 1996; Somashekar 1999), coding was tailored to the similarities of these structures across languages: lexically headed versus free relative, type of *wh*-word heading the relative clause, and the similarities of the expected response to the stimuli across languages, whether the subjects' imitation had matched the target or not. The coding also had to reflect the differences across languages, that is, their language-specific characteristics, for example, information of the relativized position was needed in French but not in English or Tulu; specific morphemes appear in Tulu but not in the other two languages (see Blume et al. in this volume, for a detailed explanation). The data complexity here is not only morphological complexity; it is relational complexity—that is, relation of discrete parts of the child's form to other discrete parts, and relation of each to the parts of the stimulus form.

Language-Variety Differences

Research with multilingual populations frequently involves working with better-known Indo-European languages, as well as lesser-studied languages such as Haitian Creole, Yiddish, and Quechua. This type of research, just as do cross-linguistic studies, needs researchers to include in addition detailed and calibrated information on the language variety, so that cross-linguistic development can be compared.

Language Switching

Multilingual populations may also switch back and forth between languages in a single transcript or within utterances (i.e., code-switching/mixing data). For example, in an experimental study attempting to measure adult code-switching, Kang, Martohardjono, and Lust (forthcoming) asked English-Chinese multilinguals to switch back and forth between their two languages. Participants were given various topics to talk about for two minutes each and were instructed to switch from one language to another upon hearing a beep. These beeps occurred every 30 seconds. Markup was developed to identify the languages at multiple levels (e.g., lexical, morphological, syntactic), in order to examine the types of switches made (e.g., do participants switch faster when they switch functional items, such as discourse connectors or content words?). This requires any coding tool either to switch easily between the markups appropriate for each language or to allow for several coding fields in each screen.



Utterance

所以这些是我。。更比较喜欢在呆Cornell的summer的夏天。然后我在工作上面的话，其实还是比较X的，没有什么，太多的，有趣的东西

Language [clear] ○ English ● Mandarin ○ Interviewer

Switch [clear] ○ Yes ● No

Pause Length [clear]

Filler [clear] ☐ Word ☐ Laugh

Code Mix [clear] ● English within Mandarin ○ Mandarin within English

Continue through switch [clear] ○ English into Mandarin ○ Mandarin into English

SUBJECT	really enjoyful	7/14
INTERVIEWER	[BEEP]	4/14
SUBJECT	所以这些是我。。更比较喜欢在呆Cornell的summer的夏天。然后我在工作上面的话，其实还是比较X的，没有什么，太多的，有趣的东西	5/14
INTERVIEWER	[BEEP]	4/14
SUBJECT	Uhm	7/14
SUBJECT	I, in-	7/14
SUBJECT	the work that I did was	7/14
SUBJECT	not very interesting, I guess, it's	7/14
SUBJECT	uhm basically, doing sale	8/14
SUBJECT	sheets, X data	7/14
SUBJECT	for them	7/14
INTERVIEWER	[BEEP]	4/14
INTERVIEWER	ok, that's it	4/14

Showing: 51 to 70 of 70

Figure 10.1
Markup created in the Data Transcription and Analysis Tool (DTA).

Working with code-switching data may imply the need to code for elements linked to language processing. For example, this experimental study focused on both fluency (defined as the time taken to switch from one language into the other after the beep) and productivity (defined as the number of words produced within two minutes), besides the types of switches. Figure 10.1 shows some of this markup created in the Data Transcription and Analysis Tool (DTA).

Multimodal Data Markup

Another set of issues pertains to the modality in which languages are expressed as well as the status and information of the language(s) under investigation. While many studies have focused on the acquisition and use of two *spoken* languages, individuals who acquire more than one sign language and those who acquire *both* spoken and signed languages are also multilingual. The transcription and analysis of sign languages present specific challenges: They do not benefit from standard orthography, and no notation system for them is currently standard (Baker, van den Bogaerde, and Woll 2008).[10] The simultaneous use of different channels of speech production—the hands and the face—complicate the accurate representation of the different components of the utterance and may have modality-specific effects in the context of interactions (Morgan, Barrière, and Woll 2006). With respect to multilingual children's acquisition of both a spoken and a sign language, research shows that "Deaf children in such a multilingual situation often produce utterances in which both the manual and vocal channels are used simultaneously" (Baker, van den Bogaerde, and Woll 2008, 20). The meanings expressed through each distinct channel may be separate or may combine, in which case transcribing the two independently from each other may not provide an accurate meaning of the full proposition (Baker, van den Bogaerde, and Woll 2008). Ultimately, data will need to be shared across researchers who work with both spoken and signed languages.

Experimental Data

As we saw in the case of the code-switching study, experimental data, for multilinguals as well as for monolinguals, require specific markup, depending on the method used. Given the current variation regarding both designs of experiments and coding systems by research teams, one needs to be able to calibrate results of different extensive markup systems indicating, for instance, the type of response (e.g., looking, pointing, moving props and toys, speaking), the timing of exposure to relevant stimuli (e.g., the point at which a child hears verbal stimuli when presented with visual stimuli in a picture- or video-matching task), and the data source (total looking time versus first long gaze in an Intermodal Preferential Looking Paradigm).

Linking Data to Metadata

As we hope to have shown in our discussion of the complexities of multilingual data, any study of language development or use must link data to rich metadata; for example, the code-switching study above looked at how attitudes toward code-switching and frequency of code-switching influenced its productivity and its fluency. Having each participant's metadata on hand in the same database is, therefore, critical for several reasons.

Design of Any Primary Research Tool Appropriate to Achieve the Vision of LLOD

It is obvious for the linguistic community working on multilingual acquisition and use that sharing data in an LLOD approach is essential to the progress of the field, since it enables us to replicate studies[11] and make full use of or reanalyze data that already exists. As we have shown, sharing Open Data would be most advantageous in terms of increasing sample sizes, allowing the identification of comparable populations, and allowing for meta-analyses.

Being able to share these data requires us to (1) standardize assessment tools as well as questionnaires, (2) capture metadata and data in efficient ways and in a design that is informed by past research, (3) link across projects and datasets, (4) allow for the capacity to query fields and relations among fields, and (5) at the same time allow for enough flexibility to capture the large diversity and richness of multilingual data.

Below, we discuss the capabilities of the Data Transcription and Analysis Tool (DTA)—but only briefly, since this tool is discussed in more detail in Blume et al. (this volume)—as an example of what is entailed in transforming any primary research tool to allow for the LLOD vision in multilingualism. The DTA tool is a primary research web application created mainly for the study of monolingual and multilingual language acquisition; it features a powerful relational database that handles both experimental and naturalistic data.

The DTA tool structures both the metadata documentation and the data creation process. It allows researchers to use built-in labels or to create project-specific labels (*codings*) to code their data, which in turn enables them to perform multiple types of analyses on their own data as well as to link data across projects.

A tool such as the DTA tool achieves requirements 2, 3, and 4 above, thus enabling researchers to share experimental (and natural speech) data so that people with varying types of expertise can reuse and repurpose them. Since the metadata and markup are so clear and specific, it becomes easy for new researchers to find all the details of a study in one place and then use that information to critique, reanalyze, and, if desired, repurpose the data.

However, the data creation process still requires many hours of dedicated and detailed work by individual researchers, since little is automated. With large sets of data, this process can take many years, so collaboration would be welcomed with other tools that have already achieved some level of automation or more efficient ways to speed up data creation (e.g., the CHILDES' CLAN[12] system or the LENA system[13]).

In terms of requirement number 4, although the tool is extremely flexible, dealing with the type of data we have described above entails some adjustments—some easier than others, but all possible. For example, capturing multimodal data would require us to display videos in the coding screen and not merely on the transcription screen. This is easily achieved and it would benefit all forms of language coding. Creating specific codes for sign language is now possible, but linking video and transcript/code is very time-consuming on the system currently available. Another challenge is that of language switch. At this point, there is no efficient way to tag the language of every word in an utterance. While this can be achieved by breaking the utterance word by word and tagging each word, this clearly could be better resolved by some automated process that may be available elsewhere.

To achieve Open Data, any tool needs to be able to speak to other tools and databases, and this bring us back to our first and major challenge. Having data that are really comparable across projects will never be achieved until we solve the standardization issues on metadata collection and presentation in requirement 1.

In sum, having an LLOD perspective and then acquiring and using any primary research tool that would aid researchers to achieve linking of their data in the study of multilingualism would require a cyberinfrastructure to support collaborative cross-linguistic research, calibration of complex multilingual markup systems, and the capacity to store, link, and search through extensive metadata.

Acknowledgments

Funding for Dr. Barrière was provided by NSF, USA/BCS#1251828 and 1251707 awarded to I. Barrière and G. Legendre; ESRC, UK; PSC-CUNY. Dr. Barrière would also like to acknowledge her collaborators: Katsiaryna Aharodnik (Graduate Center City University of New York, USA), Jennifer Culbertson (University of Edinburgh, Scotland), Guetjens (Prince) Fleurio (ENARTS, Port-au-Prince, Haiti), Nayeli Gonzalez-Gomez (Oxford University, Brooks, UK), Lisa Hsin (Harvard University, Boston, USA), Blandine Joseph (Long Island University, Brooklyn, USA), Sarah Kresh (Graduate Center City University of New York, USA), Géraldine Legendre (Johns Hopkins University, Baltimore, USA), Gary Morgan (City University, London, UK), Thierry Nazzi (University of Paris V & Centre National de la Recherche Scientifique, France), Bencie Woll (University College, London, UK), and Erin Zaroukian (US Army Research Laboratory, USA).

The authors would like to thank their colleagues at the VCLA for their input and discussion of these matters.

Notes

1. Multilingual speakers can be classified in many different ways. This classification intends to summarize and simplify on major life circumstances that may determine the speaker's level of competence and use.

2. For an example of an extensive questionnaire (78 questions, 42 pages long), see Blume and Lust's (2017). supplemental site: http://pubs.apa.org/books/supp/blume/?_ga=1.998898.2130472459.1479745044.

3. Multiple independently created questionnaires are available. One important task would be to compare them and decide which questions truly help researchers classify speakers so that a standard "minimal level" questionnaire can be created that also enables independent researchers to add questions as needed for their particular studies.

4. Pease-Álvarez, Hakuta, and Bayley (1996) also found discrepancies between children's linguistic abilities and their linguistic history.

5. https://www.iris-database.org/iris/app/home/index.

6. https://webdta.clal.cornell.edu/.

7. https://tla.mpi.nl/.

8. https://osf.io/.

9. http://ewave-atlas.org/.

10. ASL SignBank! is now being developed at the University of Connecticut by Diane Lillo-Martin and the members of the Sign Linguistics & Language Acquisition Lab.

11. This is being done for psychological studies in the Estimating the Reproducibility of Psychological Science project (https://osf.io/ezcuj/wiki/home/) and for second language acquisition by the Effects of Attention to Form on Second Language Comprehension: A Multi-Site Replication Study (https://osf.io/tvuer/), both hosted inside OSF.

12. http://dali.talkbank.org/clan/.
13. https://www.lena.org.

References

Alcock, K. J., K. Rimba, P. Holding, P. Kitsao-Wekulo, A. Abubakar, and C. R. J. C. Newton. 2015. "Developmental Inventories Using Illiterate Parents as Informants: Communicative Development Inventory (CDI) Adaptation for Two Kenyan Languages." *Journal of Child Language* 42 (4): 763–785.

Armon-Lotem, Sharon, Susan Joffe, Hadar Abutbul-Oz, Carmit Altman, and Joel Walters, J. 2014. "Language Exposure, Ethnolinguistic Identity and Attitudes in the Acquisition of Hebrew as a Second Language among Bilingual Preschool Children from Russian and English-Speaking Backgrounds." In *Input and Experience in Bilingual Development*, edited by Theres Grüter and Johanne Paradis, 77–98. Philadelphia: John Benjamins.

Austin, Jennifer B., María Blume, and Liliana Sánchez. 2015. *Bilingualism in the Spanish-Speaking World*. New York: Cambridge University Press.

Baker, Anne, Beppie van den Bogaerde, and Bencie Woll. 2008. "Methods and Procedures in Sign Language Acquisition Studies." In *Sign Language Acquisition,* edited by Anne Baker and Bencie Woll, 1–50. Philadelphia: John Benjamins.

Barac, Raluca, Ellen Bialystok, Dina C. Castro, and Marta Sanchez. 2014. "The Cognitive Development of Young Dual Language Learners: A Critical Review." *Early Childhood Research Quarterly* 29:699–714.

Barrière, Isabelle. 2010. "The Vitality of Yiddish among Hasidic Infants and Toddlers in a Low SES Preschool in Brooklyn." In *Yiddish—a Jewish National Language at 100*. Proceedings of Czernowitz Yiddish Language 2008 International Centenary Conference, edited by Wolf Moskovich, 170–196. Jerusalem-Kyiv, Hebrew University of Jerusalem.

Barrière, Isabelle. 2014. "Assessment of Language Abilities." In *Encyclopedia of Language Development*, edited by Patricia J. Brooks and Vera Kempe, 21–25. Sage.

Barrière, Isabelle, Sarah Kresh, Victoria Fay, Erika Lanham, Claribel Polanco, Stephanie Rauber, Jenice Robertson, et al. 2015. "The Contribution of Language-Specific Characteristics to Spanish-English Bilingual Preschoolers' Comprehension of Subject–Verb Agreement." Paper presented at the Tenth International Symposium on Bilingualism, Rutgers University, New Brunswick, NJ, May 20–24.

Barrière, Isabelle, Sarah Kresh, Katsiaryna Aharodnik, Géraldine Legendre, and Thierry Nazzi. 2016. "The Comprehension of 3rd Person Subject–Verb Agreement by Low SES NYC English-Speaking Preschoolers Acquiring Different Varieties of English: A Mutidimensional Approach." Paper presented at the Third Formal Ways of Analyzing Variation Workshop, Graduate Center, CUNY, New York, May 18–19.

Bialystok, Ellen. 2007. "Acquisition of Literacy in Bilingual Children: A Framework for Research." *Language Learning* 57 (Suppl. 1): 45–77.

Blume, María, and Barbara C. Lust. 2017. *Research Methods in Language Acquisition: Principles, Procedures, and Practices*. Washington, DC: American Psychological Association and De Gruyter Mouton.

Canale, Michael. 1983. "From Communicative Competence to Communicative Language Pedagogy." In *Language and Communication*, edited by Jack C. Richards and Richard W. Schmidt, 2–27. London: Longman.

Canale, Michael, and Merrill Swain. 1980. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing." *Applied Linguistics* 1:1–47.

De Houwer, Annick. 2009. *Bilingual First Language Acquisition*. Bristol/Buffalo/Toronto: Multilingual Matters.

Dunn, Lloyd M., and Douglas M. Dunn. 2007. *Peabody Picture Vocabulary Test*, 4th ed. (PPVT-4). Pearson Education.

Espinosa, Linda M., and Eugene García. 2012. "Developmental Assessments of Young Dual Language Learners with a Focus on Kindergarten Entry Assessments: Implications for State Policies." Working paper # 1, Center for Early Care and Education Research-Dual Language Learners (CECER-DLL). Chapel Hill: University of North Carolina, Frank Porter Graham Child Development Institute, 1–16.

Esquinca, Alberto, David Yaden, and Robert Rueda. 2005. "Current Language Proficiency Tests and Their Implications for Preschool English Language Learners." In *Proceedings of the Fourth International Symposium on Bilingualism*, edited by James Cohen, Kara T. McAlister, Kellie Rolstad, and Jeff MacSwan, 674–680. Somerville, MA: Cascadilla Press.

Fishman, Joshua A. 1965. "Who Speaks What Language to Whom and When?" *La Linguistique* 2:67–68.

Flege, James Emile, Ian R. A. MacKay, and Thorsten Piske. 2002. "Assessing Bilingual Dominance." *Applied Psycholinguistics* 23:567–598

Flynn, Suzanne, and Barbara C. Lust. 1981. "Acquisition of Relative Clauses: Developmental Changes in Their Heads." In *Cornell Working Papers in Linguistics*, 2 (Spring), edited by Wayne Harbert and Julia Herschensohn, 33–45. Ithaca, NY: Department of Modern Languages and Linguistics, Cornell University.

Foley, Claire. 1996. "Knowledge of the Syntax of Operators in the Initial State: The Acquisition of Relative Clauses in French and English." PhD diss., Cornell University.

Gambino, Christine P., Yesenia D. Acosta, and Elizabeth M. Grieco. 2014. "English-Speaking Ability of the Foreign-Born Population in the United States: 2012." American Community Survey Reports. U.S. Census Bureau. Accessed November 10, 2017. https://www.census.gov/library/publications/2014/acs/acs-26.html.

García, Ofelia, Zeena Zakharia, and Bahra Otcu. 2013. *Bilingual Community Education for American Children: Beyond Heritage Languages in a Global City*. Bristol, UK: Multilingual Matters.

Gathercole, Virginia C. M. 2010. "Bilingual Children: Language and Assessment Issues for Educators." In *International Handbook of Psychology in Education,* edited by Karen Littleton, Claire Wood, and Judith Kleine Staarman, 713–748. Bingley, UK: Emerald Group.

Genesee, Fred. 1989. "Early Bilingual Development: One Language or Two." *Journal of Child Language* 16:161–179. Reproduced in *The Bilingualism Reader*, edited by Li Wei, 327–343. London: Routledge.

Genesee, Fred, Elena Nicoladis, and Johanne Paradis. 1995. "Language Differentiation in Early Bilingual Development." *Journal of Child Language* 22:611–631.

Grosjean, François. 1989. "Neurolinguists, Beware! The Bilingual Is Not Two Monolinguals in One Person." *Brain and Language* 36 (1): 3–15.

Grosjean, François. 2008, 2011. *Studying Bilinguals*. Oxford: Oxford University Press.

Grosjean, François. 2010. *Bilingual: Life and Reality*. Cambridge: Harvard University Press.

Gutierrez-Clellen, Vera, and Jacqueline Kreiter. 2003. "Understanding Child Bilingual Acquisition Using Parent and Teacher Reports." *Applied Psycholinguistics* 24:267–288.

Hamers, Josiane F., and Michel H. A. Blanc. 1989. *Bilinguality and Bilingualism*. Cambridge: Cambridge University Press.

Kang, Carissa, Gita Martohardjono, and Barbara C. Lust. (unpublished manuscript). "Underlying Cognitive Mechanism for Code-switching Differs across Bilinguals."

Lambert, Wallace E. 1955. "Measurement of the Linguistic Dominance in Bilinguals." *Journal of Abnormal and Social Psychology* 50:197–200.

Lust, Barbara C., Suzanne Flynn, María Blume, Seong Won Park, Carissa Kang, Sujin Yang, and Ah-Young Kim. 2014. "Assessing Child Bilingualism: Direct Assessment of Bilingual Syntax Amends Caretaker Report." *International Journal of Bilingualism* 20 (2): 153–172.

Mackey, William. 2012. "Bilingualism in North America." In *Handbook of Bilingualism and Multilingualism*, edited by Tej K. Bhatia and William C. Ritchie, 707–724. Oxford: Blackwell.

McCabe, Alyssa, Catherine S. Tamis-LeMonda, Mark H. Bornstein, Carolyn Brockmeyer Cates, Roberta Golinkoff, Alison Wishard Guerra, Kathy Hirsh-Pasek, et al. 2013. "Multilingual Children beyond Myths and towards Best Practices." *Social Policy Report* 27 (4): 1–37.

Miller, Karen, and Cristina Schmitt. 2010. "Effects of Variable Input in the Acquisition of Plural in Two Dialects of Spanish." *Lingua* 120 (5): 1178–1193.

Morgan, Gary, Isabelle Barrière, and Bencie Woll. 2006. "The Influence of Typology and Modality in the Acquisition of Verbal Agreement in British Sign Language." *First Language* 26 (1): 19–43.

Otheguy, Ricardo, and Ana Celia Zentella. 2012. *Spanish in New York: Language Contact, Dialect Leveling and Structural Continuity*. Oxford: Oxford University Press.

Paradis, Johanne, Kristyn Emmerzael, and Tamara Sorenson Duncan. 2010. "Assessment of English Language Learners: Using Parent Report on First Language Development." *Journal of Communication Disorders* 43:474–497.

Pease-Álvarez, Lucinda, Kenji Hakuta, and Robert Bayley. 1996. "Spanish Proficiency and Language Use in California's Mexicano Community." *Southwest Journal of Linguistics* 15:137–51.

Peña, Elisabeth D. 2007. "Lost in Translation: Methodological Considerations in Cross-Cultural Research." *Child Development* 78 (4): 1255–1264.

Sánchez, Liliana. 2015. "Crosslinguistic Influences, Functional Interference, Feature Reassembly and Functional Convergence in Quechua–Spanish Bilingualism." In *The Acquisition of Spanish as a Second Language: Data from Understudied Language Pairings*, edited by S. Perpiñan and T. Judy, 19–48. Amsterdam: John Benjamins.

Somashekar, Shamitha. 1999. "Developmental Trends in the Acquisition of Relative Clauses: Cross-Linguistic Experimental Study of Tulu." PhD diss., Cornell University.

Special Eurobarometer 386. 2012. "Europeans and Their Languages Report." European Commission, Brussels. Accessed November 10, 2017. ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_386_en.pdf.

Squires, Jane, Diane Bricker, and LaWanda Potter. 1997. "Revision of Parent-Completed Developmental Screening Tool: Ages and Stages Questionnaire." *Journal of Pediatric Psychology* 22:313–328.

Thomas, Margaret. 1994. "Assessment of L2 Proficiency in Second Language Acquisition Research." *Language Learning* 44 (2): 307–336.

Thordaardottir, Elin, and Susan Ellis Weismer. 1996. "Language Assessment via Parent Report: Development of Screening Instrument for Icelandic Children." *First Language* 16:265–285.

11

Research Libraries as Partners in Ensuring the Sustainability of E-science Collaborations

Oya Y. Rieger

E-science: Fourth Paradigm, Linked Data, and Research Libraries

Over the past two decades, advances in information and communication technologies have ushered in new modes of knowledge creation, dissemination, sharing, and enquiry. These affordances, combined with the vision of global collaborations, have stimulated the development of a range of open science principles. The vision of an open and robust information infrastructure is to facilitate the broad dissemination of research outputs of all types—including research data—to allow their use, refinement, nullification, and reuse. Modern scientific instruments enabled the collection and analysis of large quantities of data, and it was almost a decade ago that Jim Gray coined the term "Fourth Paradigm" to signal the promise of data-driven scientific discovery (Hey, Tansley, and Tolle, 2007). He argued that in addition to observational, theoretical, and computational methods, data will play a significant role in advancing science. As a computer scientist, Gray emphasized the importance of developing new ways to organize, retrieve, validate, link, authenticate, and interpret data. He characterized the data-driven research life cycle with interlinked stages of data acquisition, visualization, analysis, data mining, dissemination, and archiving—and, most importantly, collaboration.

Many academic libraries' introduction to research data management came through Gray's vision of the Fourth Paradigm. This occurred during a time when libraries were starting to explore their role in the newly emerging digital scholarship landscape. Since then, several research libraries have expanded their services to collaborate with scientists in developing and maintaining new research and scholarly communication initiatives. Another influential development has been the emergence of public access requirements associated with governmental and private research funders for providing unrestricted access to research results that are produced as a result of their support. Academic libraries have been broadening their services to work with faculty in developing and implementing data management plans. This is a natural extension of their roles, since the core mission of research libraries has always been to curate the scholarly record and to make it both accessible and usable for current and future users. The Sloan Digital Sky Survey (SDSS)[1] illustrates the role of scientific collaborations by bringing together more than 25 worldwide institutions. As the project came to completion in 2008, the University of Chicago Library undertook a pilot project to investigate the feasibility of long-term storage and archiving of its data, amounting to nearly 100 terabytes (Kern et al. 2010). Library professionals contributed to the research data stewardship process through their expertise in data collection and organization, metadata creation, interoperability standards implementation, user support, and preservation. Because most of the e-science projects are supported by one-time research funds, a principal role for academic libraries is ensuring the sustainability of these initiatives beyond the project duration, so as to increase their impact and influence.

About the same time that "Fourth Paradigm" was coined, Tim Berners-Lee came up with the concept of Linked Data to promote structured data that can be interlinked to make data more accessible, usable, and shareable. Creating a ubiquitous, comprehensive, and linked research data environment is the ultimate vision, and achieving it necessitates the development of a seamless network of content, technologies, policies, expertise, and practices. There is an ongoing need for tools and methodologies that enable data processing, analysis, and visualization, along with the ability to *link* various related scholarly outputs. As we strive toward this goal, it is critical that we view each scholarly organization as an enterprise that needs to be maintained, improved, assessed, and promoted over time. Given the wide range of technical and functional requirements, the building of open information infrastructures requires bringing together the expertise of scientists, specialists, technologies, and librarians alike.

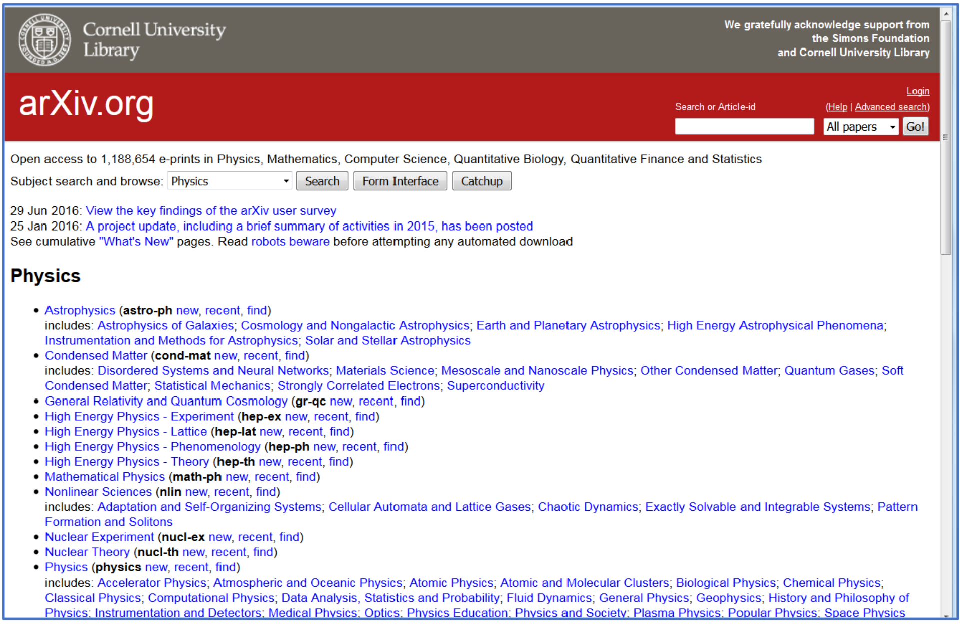
Current changes in technology and research requirements both opens up new opportunities and presents challenges in the way that research is produced, shared, preserved, and archived for future generations. To facilitate research processes from investigation to the dissemination stage, many research libraries are now extending their programs to support new modes of publishing and to facilitate the exploration of novel methodologies in performing digital scholarship. As we are engaged in Open Data initiatives, it is critical that we consider long-term development and management issues upstream as a component of an enduring service infrastructure. Simply put, sustainability is the capacity to endure; it entails long-term stewardship for responsible as well as innovative management of resource use. At the heart of this concept is the ability to secure resources (technologies, expertise, policies, visions, standards, and so on) needed to protect and enhance the value of a service based on a user community's requirements and vision.

The term "infrastructure" refers to structures, systems, and facilities that provide baseline services to a community or region, such as roads, bridges, and water systems. In the case of e-science, an infrastructure entails digital facilities, tools, services, policies, structures, and best practices that enable the creation and maintenance of a reliable network of requisite services. Basic examples include wireless networks, mass storage devices, data analysis and visualization tools, data and code repositories—to name just a few. In addition, consideration must be taken of social and organizational practices and norms, such as data access and the sharing ethos of various disciplines.

E-science initiatives require a shared infrastructure that should be seen as a public good needing to be sustained overtime (Rieger 2013). The remainder of this article, then, will focus on a case study to illustrate how libraries are involved in supporting the creation of a sustainable e-science infrastructure. Since this article was written, arXiv moved from Cornell University Library to Cornell Computing and Information Science. This transition was a natural stage in the evolution of arXiv, required for optimum service delivery and infrastructure sustainability.[2]

arXiv: E-science as Scholarly Enterprise

Started in August 1991 by Paul Ginsparg, arXiv.org is internationally acknowledged as a pioneering digital archive and open-access distribution service for research articles (see figures 11.1–11.3). This e-print repository, which moved to the Cornell University 10 years later, has transformed the scholarly communication infrastructure of multiple fields of physics and continues to play an increasingly prominent role in mathematics, computer science, quantitative biology, quantitative finance, and statistics. As of August 2016, arXiv


Cornell University Library
We gratefully acknowledge support from the Simons Foundation and Cornell University Library
arXiv.org
Login
Search or Article-id
(Help | Advanced search)
All papers
Go!
Physics
Search
Form Interface
Catchup

Open access to 1,188,654 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

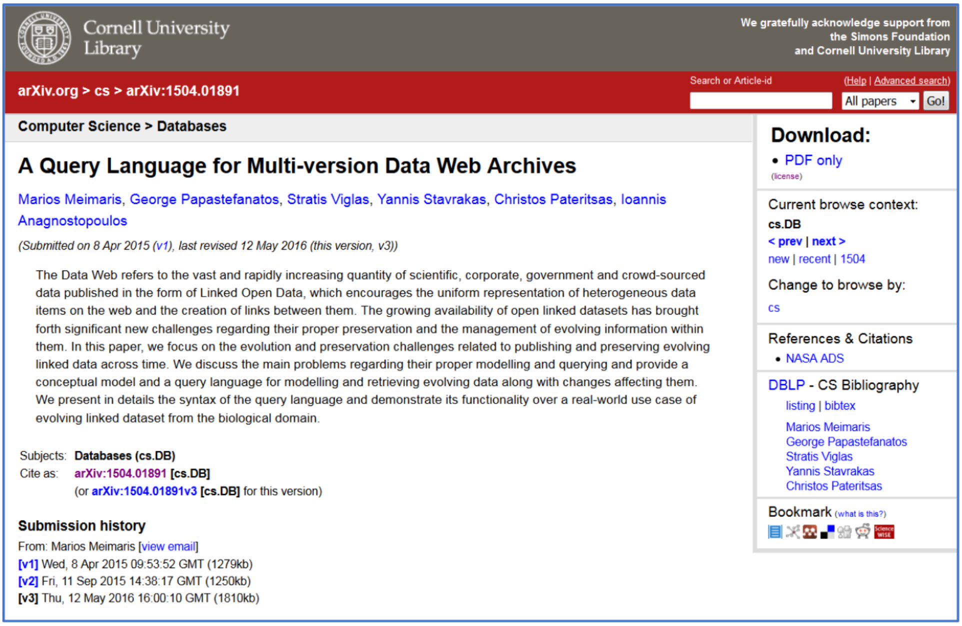
Subject search and browse: Physics

29 Jun 2016: View the key findings of the arXiv user survey
25 Jan 2016: A project update, including a brief summary of activities in 2015, has been posted
See cumulative "What's New" pages. Read robots beware before attempting any automated download

Physics

- Astrophysics (**astro-ph** new, recent, find)
  includes: Astrophysics of Galaxies; Cosmology and Nongalactic Astrophysics; Earth and Planetary Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics
- Condensed Matter (**cond-mat** new, recent, find)
  includes: Disordered Systems and Neural Networks; Materials Science; Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity
- General Relativity and Quantum Cosmology (**gr-qc** new, recent, find)
- High Energy Physics - Experiment (**hep-ex** new, recent, find)
- High Energy Physics - Lattice (**hep-lat** new, recent, find)
- High Energy Physics - Phenomenology (**hep-ph** new, recent, find)
- High Energy Physics - Theory (**hep-th** new, recent, find)
- Mathematical Physics (**math-ph** new, recent, find)
- Nonlinear Sciences (**nlin** new, recent, find)
  includes: Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons
- Nuclear Experiment (**nucl-ex** new, recent, find)
- Nuclear Theory (**nucl-th** new, recent, find)
- Physics (**physics** new, recent, find)
  includes: Accelerator Physics; Atmospheric and Oceanic Physics; Atomic Physics; Atomic and Molecular Clusters; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Fluid Dynamics; General Physics; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Physics and Society; Plasma Physics; Popular Physics; Space Physics

Figure 11.1
arXiv Homepage.



Computer Science > Databases

A Query Language for Multi-version Data Web Archives

Marios Meimaris, George Papastefanatos, Stratis Viglas, Yannis Stavrakas, Christos Pateritsas, Ioannis Anagnostopoulos

(Submitted on 8 Apr 2015 (v1), last revised 12 May 2016 (this version, v3))

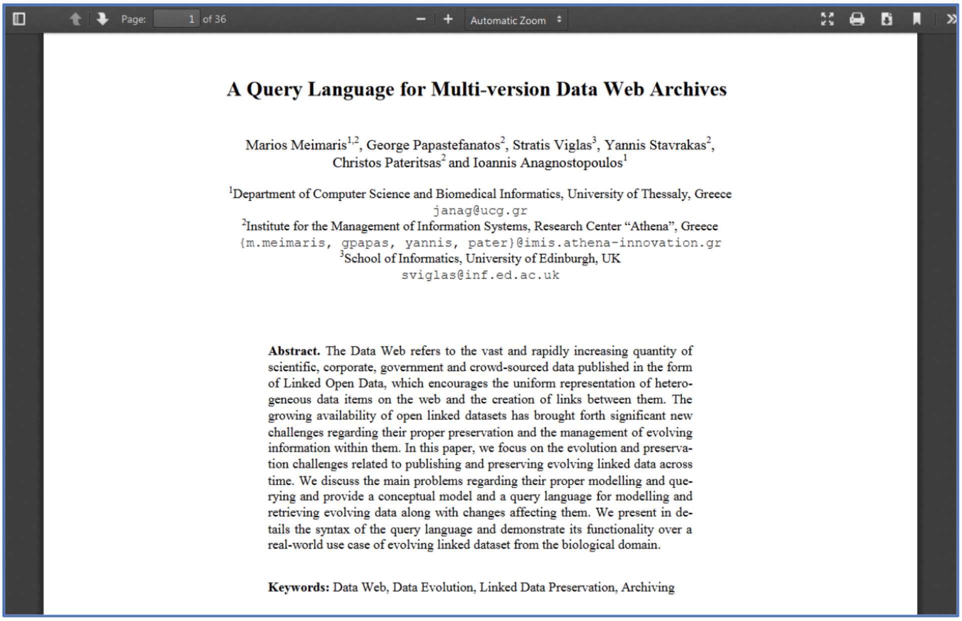
The Data Web refers to the vast and rapidly increasing quantity of scientific, corporate, government and crowd-sourced data published in the form of Linked Open Data, which encourages the uniform representation of heterogeneous data items on the web and the creation of links between them. The growing availability of open linked datasets has brought forth significant new challenges regarding their proper preservation and the management of evolving information within them. In this paper, we focus on the evolution and preservation challenges related to publishing and preserving evolving linked data across time. We discuss the main problems regarding their proper modelling and querying and provide a conceptual model and a query language for modelling and retrieving evolving data along with changes affecting them. We present in details the syntax of the query language and demonstrate its functionality over a real-world use case of evolving linked dataset from the biological domain.

Subjects: **Databases (cs.DB)**
Cite as: **arXiv:1504.01891 [cs.DB]**
(or **arXiv:1504.01891v3 [cs.DB]** for this version)

Submission history

From: Marios Meimaris [view email]
[v1] Wed, 8 Apr 2015 09:53:52 GMT (1279kb)
[v2] Fri, 11 Sep 2015 14:38:17 GMT (1250kb)
[v3] Thu, 12 May 2016 16:00:10 GMT (1810kb)

Figure 11.2
arXiv Abstract Page.



A Query Language for Multi-version Data Web Archives

Marios Meimaris[1,2], George Papastefanatos[2], Stratis Viglas[3], Yannis Stavrakas[2],
Christos Pateritsas[2] and Ioannis Anagnostopoulos[1]

[1]Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece
janag@ucg.gr
[2]Institute for the Management of Information Systems, Research Center “Athena”, Greece
{m.meimaris, gpapas, yannis, pater}@imis.athena-innovation.gr
[3]School of Informatics, University of Edinburgh, UK
sviglas@inf.ed.ac.uk

Abstract. The Data Web refers to the vast and rapidly increasing quantity of scientific, corporate, government and crowd-sourced data published in the form of Linked Open Data, which encourages the uniform representation of heterogeneous data items on the web and the creation of links between them. The growing availability of open linked datasets has brought forth significant new challenges regarding their proper preservation and the management of evolving information within them. In this paper, we focus on the evolution and preservation challenges related to publishing and preserving evolving linked data across time. We discuss the main problems regarding their proper modelling and querying and provide a conceptual model and a query language for modelling and retrieving evolving data along with changes affecting them. We present in details the syntax of the query language and demonstrate its functionality over a real-world use case of evolving linked dataset from the biological domain.

Keywords: Data Web, Data Evolution, Linked Data Preservation, Archiving

Figure 11.3
arXiv Paper View.

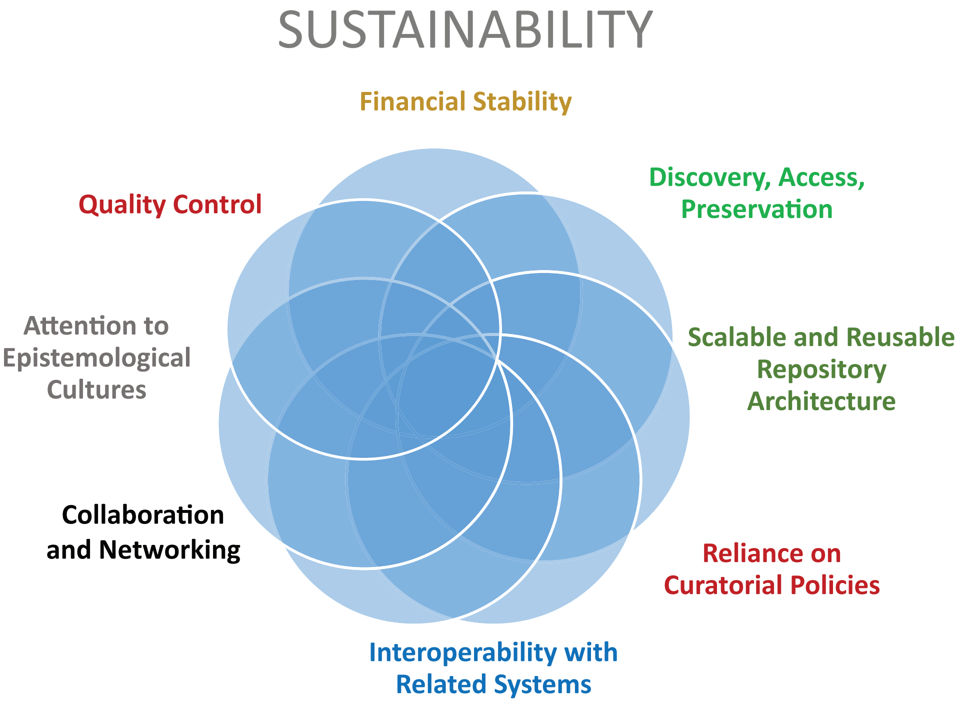

SUSTAINABILITY
Financial Stability
Discovery, Access, Preservation
Scalable and Reusable Repository Architecture
Reliance on Curatorial Policies
Interoperability with Related Systems
Collaboration and Networking
Attention to Epistemological Cultures
Quality Control

Figure 11.4
Sustainability Wheel.

included more than 1.2 million e-prints; arXiv's operating costs for 2016 were projected to be approximately $1.2 million, including salaries of eight full-time employees, server maintenance, and networking.

Since 2010, Cornell's sustainability planning initiative has aimed to reduce arXiv's financial burden and dependence on a single institution, instead creating a broad-based, community-supported resource. This sustainability initiative strives to strengthen arXiv's technical, service, financial, and policy infrastructure (figure 11.4). As a sociotechnical system, arXiv consists not only of numerous technical systems and standards but also of consistent practices and policies that are deeply embedded in the disciplines that arXiv serves. The sustainability planning process for arXiv involved building a community along with a governance system to diversify revenues.

The following section outlines five sustainability principles for e-science initiatives, based on Cornell University's experience in running arXiv (Rieger 2011).

Deep Integration into the Scholarly Community

Disciplinary characteristics, work practices, and conventions of academia all play important roles in researchers' assessment and appropriation of information and communication technologies. The information and communication technology integration that characterizes many disciplinary communities often mirrors various underlying differences in epistemic cultures. arXiv is a scholarly communication forum informed and guided by scientists and the scientific cultures it serves. It is rooted in both the academic and the information science communities, and its services have consistently focused both on the epistemic cultures represented in its digital repository and on community needs.

Systematic gathering of information about users and their usage patterns can be highly instrumental in balancing the power and potential of information technologies with the appropriate needs and workflows. Although it is tempting to add new features, a balance must be achieved between evidence-driven improvements based on actual user input and the addition of experimental and novel functionalities. In 2016, I as the program director along with members of my arXiv team at Cornell conducted a user survey to seek input from the global user community about arXiv's current services and future directions. We were heartened to receive some 36,000 responses! When the topic was raised of adding new features to arXiv to better facilitate the goals of open science, the prevailing opinion expressed was that any such features need to be implemented extremely carefully and systematically, and without jeopardizing arXiv's core values. While many respondents took the time to suggest future enhancements or the finessing of current services, several users were strident in their opposition to any changes. Throughout all the suggestions and regardless of the topic, commenters unanimously urged vigilance when approaching any changes and cautioned against turning arXiv into a "social media"–style platform (Rieger, Steinhart, and Cooper 2016). One of the survey questions sought out opinions about permitting readers to comment on papers and recommending the ones they find valuable through an annotation and ranking feature that could be added to arXiv. Although open review is emerging as a potential technique for evaluating the scientific quality and value of papers in a transparent and collaborative way, it continues to be in an experimental mode, as the scientific community explores its pros and cons. From a technological perspective, a range of applications are currently available that support open review. However, the intriguing and "tricky" parts are much more in the sociology of science domain that involves human factors, especially those related to the reputation, fairness, power dynamics, bias, civility, and qualifications of the participants in open review. These problems are not insurmountable, yet they certainly require the careful development of policies, procedures, and workflows that can ensure a trusted and useful environment for open review and annotation.

Clearly Defined Content Policies

Although arXiv is not peer-reviewed, submissions to it are reviewed by a network of some 150 subject-based moderators to ensure the scientific quality of its content, because the papers submitted are expected to be of interest, relevance, and value. Additionally, an "endorsement" system is in place to make sure that content is relevant to current research in the specified disciplines. In the aforementioned user survey, arXiv's users were asked a series of questions regarding quality-control measures. The most important of these (ranked very important/important) were: check papers for text overlap, as in plagiarism (77%); make sure submissions are correctly classified (64%); reject papers with no scientific value (60%); and reject papers with self-plagiarism (58%). Some users would prefer that arXiv embrace a more-open peer review and/or moderation process, while others felt adamant that current controls already permit arXiv to have a freedom and speed of access that are otherwise unobtainable through traditional publishing. Overall, the feeling was expressed that quality control matters, although user comments varied greatly in relation to how arXiv could achieve these goals in actual practice. As one respondent wrote, "Judgment about quality control is a very relative issue."

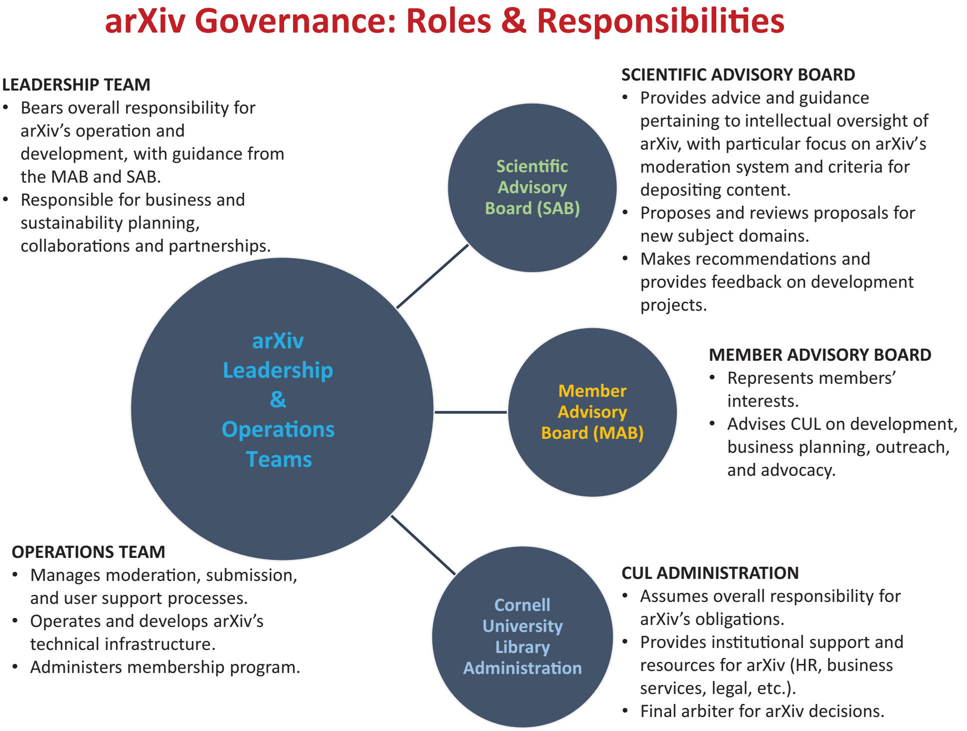
It is critical to have clearly articulated policies about the copyright status of the deposited materials, as well as conflict management processes (such as responding to concerns in regard to rejected submissions or author disputes). Cornell University's participation in the ORCID (Open Research and Contributor ID) author identifiers initiative aims to enable better author linking and to facilitate improvements in ownership claiming.

We at arXiv have adopted a measured approach to expansion, because we have found that significant organizational and administrative efforts are required to create and maintain new subject areas. Adding a new subject area involves exploring the user base and use characteristics pertaining to the subject area, establishing the necessary advisory committees, and recruiting moderators. Also, although arXiv.org is the central portal for scientific communication in some disciplines, it is neither feasible nor necessarily desirable for it to play that role in all disciplines. Although we anticipate that arXiv will become increasingly broad in its subject area coverage, we believe this development must occur in a planned, strategic manner. One of the arXiv principles is that any expansion into other subjects or disciplines must include scholarly community support, satisfy arXiv's quality standards, and take into consideration its operational capacity and financial requirements.

Clearly Defined Principles and Governance Structure

Although best practices in developing technical architectures and associated processes and policies underpin a digital repository, organizational attributes are equally important. The Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) tool, emphasizes that organizational attributes affect the performance, accountability, and sustainability of repositories.[3] The first criteria in the TRAC assessment tool are governance and organizational viability. Similarly, subject repositories must have clearly defined mandates and associated governance structures so as to reflect a commitment to the long-term stewardship of a given service. For instance, arXiv provides an open-access repository of scientific research to authors and researchers worldwide. It is a moderated scholarly communication forum informed and guided by scientists and the scientific cultures it serves. Access via arXiv.org is free to all end-users, and individual researchers can deposit their own content in arXiv for free.[4]

The general purpose of any governance is to ensure that an organization both possesses the means to envision its future and has in place management structures and processes to ensure that the envisioned plan can be implemented and sustained. Good governance is participatory, consensus oriented, accountable, transparent, responsive, efficient, equitable, and inclusive. However, it also needs to be nimble and flexible—not allowing any gridlocks or excessive groupthink. Accordingly, the arXiv membership program aims to engage those libraries and research laboratories worldwide that represent arXiv's heaviest institutional users in the service's governance. Its governance structure aims to provide a framework with well-defined roles and expectations (figure 11.5). Cornell holds the overall administrative and financial responsibility for arXiv's operation and development, with guidance from its Member Advisory Board (MAB) and its Scientific Advisory Board (SAB). Cornell manages the moderation of submissions and user support, including the development and implementation of policies and procedures, operates arXiv's technical infrastructure, assumes responsibility for archiving to ensure long-term access, oversees arXiv mirror sites, and establishes and maintains collaborations with related initiatives to improve services for the scientific community through interoperability and tool-sharing.


arXiv Governance: Roles & Responsibilities
Scientific Advisory Board (SAB)
arXiv Leadership & Operations Teams
Member Advisory Board (MAB)
Cornell University Library Administration

LEADERSHIP TEAM
- Bears overall responsibility for arXiv's operation and development, with guidance from the MAB and SAB.
- Responsible for business and sustainability planning, collaborations and partnerships.

SCIENTIFIC ADVISORY BOARD
- Provides advice and guidance pertaining to intellectual oversight of arXiv, with particular focus on arXiv's moderation system and criteria for depositing content.
- Proposes and reviews proposals for new subject domains.
- Makes recommendations and provides feedback on development projects.

MEMBER ADVISORY BOARD
- Represents members' interests.
- Advises CUL on development, business planning, outreach, and advocacy.

OPERATIONS TEAM
- Manages moderation, submission, and user support processes.
- Operates and develops arXiv's technical infrastructure.
- Administers membership program.

CUL ADMINISTRATION
- Assumes overall responsibility for arXiv's obligations.
- Provides institutional support and resources for arXiv (HR, business services, legal, etc.).
- Final arbiter for arXiv decisions.

Figure 11.5
arXiv Governance Model.

Technology Platform Stability and Innovation

The existing e-science ecology is a complex of architectures and features that are optimized to fulfill the specific needs of scientific communities. The landscape is becoming even more heterogeneous. In addition to scholarly online resources, a number of scientific social networking sites already profile local scholarly activities and host Open Data initiatives that focus on models for data curation. Several initiatives are now working on data-rich domains and share similar challenges in managing, analyzing, sharing, and archiving data. A critical component of a sustainability plan is to consider this rich context and understand how a given initiative fits within the broader framework and how we can link related information and communities. For instance, based on the findings of the arXiv user study, one important service that should expand has emerged for improving support for submitting and linking research data, code, slides, and other materials associated with papers emerged. The open text responses demonstrated considerable interest in better support for supplemental materials, although responses were divided as to whether they should be hosted by arXiv or another entity. Many respondents were supportive of integrating or linking to other services (especially GitHub), while a significant number of respondents also expressed concerns about long-term availability and “link rot” for content not hosted within arXiv. Some interest was even expressed in including the data that underly figures in arXiv papers.

Among the critical roles of repositories such as arXiv is facilitating the preservation function. Digital preservation (a term used interchangeably with “archiving”) refers to a range of managed activities to support the long-term maintenance of digital content, thereby ensuring that digital objects are usable. However, ensuring such access over time involves more than bitstream preservation.[5] The effort must provide continued access to digital content through various delivery methods, since “preserving access” entails protecting the usability of a digital resource by retaining all quantities of authenticity, accuracy, and functionality that are deemed to be essential. Therefore, preservation should be seen as a life cycle activity that requires collaboration between technology providers and user communities. Scientists are seldom experienced in preserving data for long-term access. Given the growing emphasis on open science and public access requirements, research libraries have long stood at the forefront of providing data management services that include development of preservation strategies in order to protect data for long-term access and reuse. Ideally, your own data should be regularly audited to guarantee its integrity, associated with appropriate metadata to ensure its discoverability, and monitored to control access to meet privacy, licensing, and intellectual property restrictions (Heidorn 2011; Interagency Working Group on Digital Data 2009).

An inherent tension often arises between technological stability and innovation. A reliable repository needs to be fully operational to fulfill its daily production functions, processes, and procedures. Having a dependable system in place is critical in order to provide reliable access to services on which end-users can depend. Although stability and consistency are important service attributes, also essential is keeping pace with evolving user needs through research and development (R&D) projects. Given the uncertainties associated with the development and testing of new features and services, an innovation agenda needs to be carefully thought out so as to ensure that operational stability is not undermined. Ideally, complementary streams of resources should be established to support both operational and research activities. In the case of arXiv, the membership model focuses on operational costs, and funds required for adding new features and system redesign need to be generated regularly through other revenue sources and grant-writing efforts.

Reliance on Business Planning Strategies

The primary purpose of a business planning process is to convey to its potential users a clear value proposition that will justify their investing in the business's services or product. Value propositions may be based on a range of characteristics, such as service features, customer support, product customization, and economical pricing, among many others. The key challenge in creating a value proposition is to address the needs of all stakeholders. For instance, in the case of arXiv, the stakeholders include scientists, libraries, research centers, societies, publishers, and funding agencies. Although such entities are likely to share common goals, each one attaches value to a specific aspect of arXiv. As an example, from the end-users' perspective, scientists' highest priority for arXiv is likely to be the robustness and reliability of its repository and access features.

Business plans offer an overall view of a given product, relevant user segments, key stakeholders, communication channels, competencies, resources, networks, collaborations, cost structures, and revenue models. In a collaborative model such as that of the arXiv membership, it is critical to clearly define and justify the pricing model as well as the budget so as to understand how revenues are being generated and spent. Maintaining, supporting, and further developing a repository involve a range of expenses, such as management, programming, system administration, curation, storage, hardware, facilities (space, furniture, networking, phone. and so on), research and training (such as attending meetings and conferences), outreach and promotion, user documentation, and administrative support. Also essential is to allow transparency and instill trust by sharing financial projections, roadmaps that include annual goals, and periodic reports to stakeholder communities to keep them informed and engaged.[6]

Linked Open Data: Innovation in Support of Sustainability

The Linguistic Linked Open Data (LLOD) workshop at the University of Chicago in 2015 (see the introduction to this volume) brought together a range of experts to discuss data standards, technologies, methodologies, and strategies to share a community vision for the design and implementation of a sustainable infrastructure in order to make large quantities of linguistic data accessible to a wide range of scientists and students. One of the goals was promoting Linked Open Data principles to rely on structured and nonproprietary open formats, use unique and persistent identifiers, link data to other related sources to provide context, provide mechanisms to link individual schemas and vocabularies, and privilege the use of unrestricted licenses. Such requirements not only enable data to be uniformly discoverable but also facilitate the long-term management and accessibility of digital assets. Therefore, the Open Data principles are also the key tenets of "research data sustainability" with the goal of supporting reuse that will build on and verify existing data, theory, and hypothesis to leverage our institutions' investment in scientific research. However, the increasing openness of science and the burgeoning data management mandates usher in a complex suite of technology, policy, and service needs. The arXiv case study illustrates the need to manage e-science initiatives such as the LLOD holistically, by taking into consideration a range of life cycle and usability issues, as well as factoring in changing patterns and modes characteristic of scholarly communication. We must consider the sustainability requirements upstream and remember that the services we are now experimenting with and creating have vital long-term implications.

Notes

1. The Sloan Digital Sky Survey (SDSS) is a major survey of galaxy clusters, performed with the use of using a dedicated optical telescope at Apache Point Observatory in New Mexico. The project was named after the Alfred P. Sloan Foundation, which contributed significant funding. Data collection began in 2000, and the available data today covers over 35% of the sky and includes photometric observations of around 500 million objects.

2. https://confluence.cornell.edu/display/arxivpub/Transition+FAQ%3A+Move+to+Cornell+Computing+and+Information+Science.

3. TRAC provides tools for the audit, assessment, and potential certification of digital repositories for determining the soundness and sustainability of digital repositories. For more information, see: http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.

4. See the arXiv Principles document on arXiv Public Wiki for the operational, editorial, economic, and governance principles: https://confluence.cornell.edu/display/arxivpub/arXiv+Public+Wiki.

5. Bitstream preservation aims to keep digital objects intact and readable. It ensures integrity of the bitstream by monitoring for corruption of data fixity and authenticity and also by protecting digital content from undocumented alteration, thereby securing data from unauthorized use while simultaneously providing media stability. Digital objects are items stored in a digital repository and, in their simplest form, consist of data, metadata, and an identifier.

6. Examples of such strategies can be found on the arXiv public Wiki: https://confluence.cornell.edu/display/arxivpub/arXiv+Public+Wiki.

References

Heidorn, P. B. 2011. "The Emerging Role of Libraries in Data Curation and E-science." *Journal of Library Administration [Internet],* October 51 (7–8): 662–72.

Hey, Tony, Stewart Tansley, and Kristin Tolle (editors). 2007. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Interagency Working Group on Digital Data. 2009. "Harnessing the Power of Digital Data for Science and Society." Report of the IWGDD to the Committee on Science of the National Science and Technology Council. http://www.nitrd.gov/about/harnessingpowerweb.pdf.

Kern, Barbara, Dean Armstrong, Charles Blair, David Farley, Kathleen Feeney, Eileen Ielmini, Elisabeth Long, et al. 2010. "The SDSS and E-science Archiving at the University of Chicago Library" (June 22, 2010). International Association of Scientific and Technological University Libraries, 31st Annual Conference. Paper 9. http://docs.lib.purdue.edu/iatul2010/conf/day2/9.

Rieger, Oya Y. 2011. "Assessing the Value of Open Access Information Systems: Making a Case for Community-Based Sustainability Models." *Journal of Library Administration* 51:485–506.

Rieger, Oya Y. 2012. "Subject and Institutional Archives: Comparing the Examples of arXiv and Cornell's Institutional Repository." *Insights* 25:103–106.

Rieger, Oya Y. 2013. "Sustainability: Scholarly Repository as an Enterprise." *Bulletin of the American Society for Information Science and Technology* October/November 2013. http://www.asis.org/Bulletin/Oct-12/OctNov12_Rieger.html.

Rieger, Oya Y., Gail Steinhart, Deborah Cooper. 2016. "arXiv@25: Key Findings of a User Survey." July 2016. http://arxiv.org/abs/1607.08212.

Contributors

Isabelle Barrière is an associate professor in the Department of Communication Sciences and Disorders at Long Island University/Brooklyn. Dr. Barrière is also the Director of Policy for Research and Education at the Yeled V'Yalda Early Childhood Center, one of the largest Head Start programs in New York City and a member of the Doctoral Faculty of CUNY Graduate Center. She completed her PhD in applied linguistics at Birkbeck College/University of London. She subsequently received training in early childhood education in the UK, in neuropsychology in France, and in cognitive science at Johns Hopkins University. Her research focuses on the acquisition of different languages (e.g., British Sign Language, and different varieties of English, French, Haitian Creole, Spanish, Russian, and Yiddish) with a focus on morphosyntax and a reliance on both experimental paradigms and corpus analyses. Her work has been published in peer-reviewed linguistics and psychology journals and has been supported by grant-giving organizations, including the UK ESRC, the New York State Department of Education, and the National Science Foundation. She is currently the Director of the NSF-funded Research Experience for Undergraduate site program *Intersection of Linguistics, Language and Culture*, which focuses on turning the linguistic and cultural heritage of minority students in New York City into a professional asset. https://orcid.org/0000-0002-5369-3790.

Nan Bernstein Ratner, Ed.D., CCC-SLP, is professor in the Department of Hearing and Speech Sciences at the University of Maryland, College Park. She is an ASHA Honors recipient and a Fellow of the American Association for the Advancement of Science, and a Board-recognized specialist in child language and language disorders. With Brian MacWhinney, Bernstein Ratner codirects the recent TalkBank initiative FluencyBank, funded by the National Institutes of Health and National Science Foundation. She publishes widely in the areas of language and fluency development and disorder. https://orcid.org/0000-0002-9947-0656.

Steven Bird is a linguist and computer scientist. He divides his time between Darwin, Australia's most culturally diverse city, and a remote Aboriginal community where he is learning to speak Kunwinjku. His language work has taken him to West Africa, South

America, Central Asia, and Melanesia. Bird has a PhD in computational linguistics and is professor in the College of Indigenous Futures, Arts and Society, at Charles Darwin University. He serves as Linguist at Nawarddeken Academy in West Arnhem, and Senior Research Scientist at the International Computer Science Institute, University of California–Berkeley.

María Blume is associate professor at Pontificia Universidad Católica del Perú, in Lima, where she also received her bachelor's and licentiature degrees, before continuing her studies at Cornell University, where she received her MA and PhD in linguistics. Her research interests include monolingual and bilingual acquisition as well as the relationship of morphosyntax with pragmatics. She is a founding member of the Virtual Center for the Study of Language Acquisition and of Grupo de Investigación en Adquisición del Lenguaje. Her current projects include the adaptation of the MacArthur-Bates Communicative Development Inventories to Peruvian Spanish and research on subject and verbal morphology development in bilingual and second-language acquisition. https://orcid.org/0000-0003-3786-8551.

Ted Caldwell is an application developer and senior software architect. His work at GORGES encompasses all aspects of web application development, including requirements analysis, database and software design, programming, project management, and system administration. With over 20 years of software development experience, Caldwell enjoys working closely with GORGES clients to help them understand their software needs and create practical, cost-effective technology solutions.

Christian Chiarcos is professor of computer science at Goethe University, Frankfurt, Germany, and heads the Applied Computational Linguistics group. He received a doctoral degree in natural language generation from the University of Potsdam, Germany. He subsequently worked at the Information Sciences Institute of the University of Southern California, before joining Goethe University. His research focuses on semantic technologies, including natural language understanding and knowledge representation. His specific interests cover computational discourse semantics, natural language processing, and language resource interoperability. As a computational linguist, Chiarcos explored the Semantic Web and Linked Data from an NLP perspective and contributed to the emergence of a community at the intersection of both areas. He was a cofounder of the Open Linguistics Working Group of Open Knowledge International, and he initiated and co-organized both the Linked Data in Linguistics workshop series and the accompanying development of the Linguistic Linked Open Data cloud.

Cristina Dye is assistant professor of language development and psycholinguistics at Newcastle University, UK. She obtained her PhD from Cornell University, where she was part of the Cornell Language Acquisition Lab and contributed in various ways to the Virtual Center for the Study of Language Acquisition. She conducted postdoctoral research at the Brain and Language Lab in the Department of Neuroscience at Georgetown University, and later joined Newcastle University, where she is a member of the Centre for Research in Linguistics and Language Sciences. Her research interests focus on the mechanisms involved in language development across monolingual, bilingual, and multilingual children as well as children with neurodevelopmental disorders. She regularly teaches courses and supervises dissertations in various areas of language development.

Suzanne Flynn, a founding member of the Virtual Center for the Study of Language Acquisition, has been a professor of linguistics and language acquisition at MIT since 1981. Her research focuses on the acquisition of various aspects of syntax by both children and adults in bilingual, second-language, and third-language contexts. More recently, she has expanded her work to focus on the neural representation of the multilingual brain and on the nature of language in individuals with early onset of Alzheimer's disease. Flynn has published extensively in journals, authored and coedited several books, and served as the cofounding editor of *Syntax: A Journal of Theoretical, Experimental and Interdisciplinary Research*. She received her MS from the University of Puerto Rico and her MA and PhD in linguistics from Cornell University. She has a clinical certification in speech and language pathology.

Claire Foley is a lecturer in linguistics at Boston College. Her academic background is in developmental psycholinguistics, and she has led experimental research groups and published in the areas of first- and second-language acquisition and methodology. Before going to Boston College, she worked as a visiting scholar in linguistics at MIT and as a faculty member at Morehead State University. Her PhD in linguistics is from Cornell University.

Nancy Ide is professor of computer science at Vassar College. She has been an active researcher in the field of computational linguistics for over 25 years and has published copiously on topics including computational lexicography, word sense disambiguation, Semantic Web implementations of linguistically annotated data, and standards for representing language resources and interlinked layers of linguistic annotations. She has been involved in the development of standards for language data since 1987, when she founded the Text Encoding Initiative. Later, she developed the XML Corpus Encoding Standard and, most recently, was the principal architect of the ISO Linguistic Annotation Framework. Ide is the co-editor-in-chief of the Springer journal *Language Resources and Evaluation* and editor of the Springer book series *Text, Speech, and Language Technology.* She is also the cofounder and president of the Association for Computational Linguistics special interest group on Annotation (ACL-SIGANN). Recently, she coedited a 1,300-page, two-volume book titled *Handbook of Linguistic Annotation*, which provides a definitive survey of state-of-the-art annotation practices and projects in the field. https://orcid.org/0000-0001-9473-3544.

Carissa Kang is a user experience researcher at Youtube Kids. She graduated from Cornell University with a PhD in developmental psychology. Her research centers on the study of bilingualism in the developing child, investigating several dimensions of bilingualism, among them cognitive, cultural, and social. In addition, pursuing her interest in children's cognitive development more broadly, and with cultural distinctions involved in bilingualism, she also works on how culture influences children's beliefs about choice.

D. Terence Langendoen is professor emeritus of linguistics at University of Arizona. He received his SB in philosophy and mathematics and a PhD in linguistics from MIT. He has held regular faculty positions in linguistics at Ohio State University, the Graduate Center of the City University of New York (CUNY), and University of Arizona; and in English at Brooklyn College of CUNY. He has held visiting appointments at the Hartford Seminary Foundation's Kennedy School of Missions, University of Pennsylvania, Rockefeller University, State University of New York at Buffalo, University of California at Santa Cruz, University of Utrecht, IBM T. J. Watson Research Center, and City University of Hong Kong. After retiring, he served as program director in linguistics at the National Science Foundation (NSF) for two years, and since then has been serving as expert in the Division of Information & Intelligent Systems of the Computer & Information Science & Engineering Directorate at NSF. Langendoen's current research interests include mathematical and computational linguistics, syntax, and semantics. He is author or coauthor of four books in linguistics, coeditor of three others, and has published numerous articles on a variety of topics in linguistics. He served as secretary-treasurer and as president of the Linguistic Society of America (LSA), directed the 1986 LSA Linguistic Institute at the CUNY Graduate Center, and was chair of the Linguistics and Language Sciences section of the American Association for the Advancement of Science (AAAS). He was editor of *Linguistics Abstracts*, published by Blackwell for 10 years, and book review editor for the Linguist List for six years. He is a fellow of the AAAS, LSA, Association for Psychological Science, and the New York Academy of Sciences, and was named Partner in Education by the Board of Education of the City of New York. https://orcid.org/0000-0002-3063-1896.

Barbara C. Lust is professor emerita at Cornell University. She studied developmental psychology, linguistics, and philosophy, as well as English literature. She received her PhD in developmental psychology from City University of New York Graduate Center after earlier studies at L'Institut des Sciences de l'Education, at the University of Geneva, Switzerland. She followed this with postdoctoral study in linguistics and philosophy at MIT before her teaching at Cornell, which involves developmental psychology and linguistics, within an interdisciplinary perspective of cognitive science. Lust's research is situated in an interdisciplinary and cross-linguistic framework, embracing the study of first, second, and multilingual language acquisition, especially in the child, and linking theoretical paradigms of linguistic inquiry to experimental methods of research. She is a founding member of the Virtual Center for the Study of Language Acquisition. She is also developing a comparative study of language in normal healthy aging, in contrast to that in early Alzheimer's disease. https://orcid.org/0000-0002-7923-7937.

Brian MacWhinney is professor of psychology, computational linguistics, and modern languages at Carnegie Mellon University. He received his PhD in psycholinguistics from the University of California at Berkeley. With Elizabeth Bates, he developed a model of first and second language processing and acquisition based on competition between item-based patterns. He and Catherine Snow cofounded the CHILDES (Child Language Data Exchange System) Project for the computational study of child language transcript data. The TalkBank Project extends these methods to language areas such as aphasiology, second language learning, TBI, Conversation Analysis, and others. MacWhinney's recent work includes studies of online learning of second language vocabulary and grammar, situationally embedded second language learning, neural network modeling of lexical development, fMRI studies of children with focal brain lesions, and ERP studies of between-language competition. He is also exploring the role of grammatical constructions in the marking of perspective shifting, the determination of linguistic forms across contrasting time frames, and the construction of mental models in scientific reasoning. https://orcid.org/0000-0002-4988-1342.

Jonathan Masci graduated from Cornell University with a BA in linguistics and a minor in cognitive science. As an undergraduate student, he focused on multilingualism and contributed to research on the relation between code-switching and executive function in bilingual children. He also helped to manage research projects in the Virtual Center for the Study of Language Acquisition's Data Transcription and Analysis (DTA) tool, aiming to integrate his personal interest in information technology with cognitive science research. Masci is currently an MPA student at New York University specializing in public policy analysis. https://orcid.org/0000-0002-5013-4707.

Steven Moran is a research assistant at the Department of Comparative Linguistics at the University of Zurich. He works on linguistic typology, Linked Data and under-resourced languages, language evolution, and quantitative approaches to language acquisition. He also conducts linguistic fieldwork in West Africa. https://orcid.org/0000-0002-3969-6549.

Antonio Pareja-Lora received a PhD in Computer Science and Artificial Intelligence from the Universidad Politécnica de Madrid (UPM). He is an associate professor (Profesor Contratado Doctor) at DSIC, *Departamento de Sistemas Informáticos y Computación* (Computer Systems and Computation Department), of the Universidad Complutense de Madrid (UCM). His research interests focus on natural language processing, linguistic and/or ontological annotation, and the ontological representation of linguistic phenomena, data categories, and relations. Within this area, he has collaborated in several European and national projects with the Ontology Engineering Group (OEG) of UPM for more than eight years. More recently, he has become a member of ATLAS as well as of the ILSA

(Implementation of Language-Driven Software and Applications) research group from UCM. Pareja-Lora is an expert in the ISO/TC 37 Committee (standardization of terminology, language resources, and linguistic annotations), and is also the Convenor of AENOR's AEN/CTN 191 (the national body committee corresponding to ISO/TC 37). In addition, he is one of the officers of ACL SIGANN (Special Interest Group for Annotation of the Association for Computational Linguistics). https://orcid.org/0000-0001-5804-4119.

James Reidy is a programmer working for Cornell University Library. Most of his work involves providing data for websites, restructuring them, and maintaining them. Asked to review the Data Transcription and Analysis (DTA) tool, to see how it might be archived by the Library, he became interested in the project and used time provided by the Library's Technology Innovation Week program to explore the DTA tool and try to export DTA tool data to CLAN. That turned out to be challenging, and he leaned about the enormity of the problems that linguists face. He recently learned how to make DTA tool data available for use with SPARQL queries (although not using the official Ontology) and hopes to pursue that in the future.

Oya Y. Rieger is arXiv program director at Cornell Computing and Information Science and a senior advisor to Ithaka S+R. She is involved in research and advisory projects that explore new roles for research libraries and the possibilities of open source software and open science. As arXiv.org's program director, she leads the operations, governance, sustainability, and strategic development of the open access preprint service. During her tenure at Cornell, her program areas included digital scholarship, collection development, digitization, preservation, user experience, scholarly publishing, learning technologies, research data management, digital humanities, and special collections. She spearheaded projects funded by the Institute of Museum and Library Studies, the Henry Luce Foundation, The Andrew W. Mellon Foundation, National Endowment for the Humanities, Simons Foundation, and Sloan Foundation to develop e-journal preservation strategies, conduct research on new media archiving, implement preservation programs in Asia, design digital curation curriculums, and create sustainability models for alternative publishing models to advance science communication. Throughout her career, Rieger has held numerous leadership positions with national and international organizations and has published widely on topics such as institutional and subject repositories, digital preservation, and developing sustainable business models for new forms of scholarly communication. At Cornell, she taught courses on visual research and scientific collaborations at the Communication and Architecture departments. Her ORCID identifier is https://orcid.org/0000-0001-6175-5157.

Gary F. Simons is chief research officer for SIL International and the executive editor of *Ethnologue*. He has contributed to the development of cyberinfrastructure for linguistics as cofounder of the Open Language Archives Community, codeveloper of the ISO 639-3 standard of three-letter identifiers for the known languages of the world, and codeveloper of linguistic markup for the Text Encoding Initiative. Simons was formerly director of Academic Computing for SIL International in which role he oversaw the development of linguistic annotation tools like IT (Interlinear Text Processor), CELLAR (Computing Environment for Linguistic, Literary, and Anthropological Computing), LinguaLinks, and the beginnings of FLEx (FieldWorks Language Explorer). He holds a PhD in general linguistics (with minor emphases in computer science and classics) from Cornell University.

Thorsten Trippel is a researcher of Computational Linguistics who joined the Computational Linguistics and General Linguistics group at the University of Tübingen, Germany, in 2010. There he works with CLARIN, a research infrastructure for the humanities and social sciences, providing archival systems for research data, search functionalities for and in research data, and tools analyzing language related research data. Within CLARIN he serves as liaison coordinator, reaching out to user groups and other research infrastructures and initiatives. He obtained a PhD at Bielefeld University, Germany, in its Department of Linguistics and Literary Studies. He has been working on curating linguistic resources, metadata development, tools for the evaluation of metadata and lexical resources, and language documentation of endangered languages. Trippel's expertise is in supporting other research groups in archiving and providing data, mediating between the technology and the users, and designing bridging technologies and material. Currently he serves as deputy chair of the German standardization committees for language resources and systems for managing terminology, mirroring the international activities of ISO/TC 37 SC 3 and SC 4. Within ISO/TC 37 SC4 he is the project leader for ISO 24622–2, which is the second part of the standard family "Language resource management—Component Metadata Infrastructure (CMDI)." He has coauthored more than 50 publications. https://orcid.org/0000-0002-7211-7393.

Kara Warburton is director of Business Development at Interverbum Technology. She has 35 years of experience in the translation industry where she has played a leading role in terminology management, particularly for large global enterprises. She holds PhD and MA degrees in terminology management and degrees in translation and education. Through her company, Termologic, Warburton provides consulting, support and training in terminology management, controlled authoring, localization and search engine optimization. She is currently also teaching terminology management and localization at University of Illinois.

Sue Ellen Wright is a professor emerita at Kent State University. In addition to her graduate work at Washington University, she also studied at the Johann Wolfgang von Goethe Universität Frankfurt/Main. She is an ATA accredited translator (German to English) with a specialty in manufacturing engineering, automotive engineering, electronics, and computer applications. Her teaching responsibilities included primarily the teaching of computer applications for translation in addition to German–English scientific/technical/medical translation as well as literary and cultural translation. Her primary research focus is on the management of technical terminology for translators, technical writers, and standardizers. A recognized international terminology expert, Wright is chair of the USA Technical Advisory Group (TAG) of ISO/TC 37, Language and Terminology Convener of TC 37/Sub-Committee 3/WG 1, responsible for preparing ISO 12620: Terminology–Computer Applications–Data Category Specifications, past Chair of TC 37/SC 3, Management of Terminology Resources, and a member of ASTM F43, Language Services. https://orcid.org/0000-0002-8533-6253.

Claus Zinn is a postdoctoral researcher and a research and development engineer in the Linguistics Department at the University of Tübingen, Germany. He received his diploma in computer science from the University of Erlangen-Nuremberg, where he also obtained his PhD (summa cum laude). At the University of Edinburgh, the German Research Institute for Artificial Intelligence, and the University of Konstanz, he worked on intelligent tutoring systems. Since 2007, Zinn has contributed to e-science related projects. At the Max-Planck Institute in Nijmegen, the Netherlands, he built the precursor of the CLARIN Virtual Language Observatory. At the University of Tübingen, he worked on metadata-related issues and data management planning. He also supports the CLARIN Language Resource Switchboard. Zinn has a rich experience in national and international project management and has coauthored over 50 publications. https://orcid.org/0000-0002-6067-5451.

Author Index

Abromeit, F., 16, 62
Aguado de Cea, G., xv, 13, 16, 153, 173
Alcock, K. J., 190
Armon-Lotem, S, 191
Atkins, D. E., ix, 151
Austin, J. B., 186

Baker, A., 190, 194–195
Baker, T., 120
Banisar, D., 2
Barac, R., 190
Barrière, I., 190–191, 194
Bellandi, A., 14
Bender, E., 21, 153
Benson, M., 44
Bentivogli, L., 28
Berman, F., xiv, 151
Berners-Lee, T., ix, 4, 7, 19, 107, 111, 120–121, 152, 172, 176, 202
Bernstein-Ratner, N., 145, 152–153, 172, 174
Bialystok, E., 187
Biber, D. S., xv
Bickel, B., 50
Bird, S., xiv–xv, xvii, 9, 31, 101, 118, 132, 152, 154, 173, 177
Bizer, C., 120
Bloomfield, L., 26
Blume, M., ix, xiii–xv, xvii, 132, 152–156, 158–159, 164–168, 173, 175–177, 186, 193, 196
Bond, F., 28
Borgman, C. L., ix, xi–xiii, 151–152, 155
Bray, T., 30
Broeder, D., 102
Brown, C., 58
Brown, R., 27, 29, 44, 143
Brummer, M., 53
Burchardt, A., 14
Byrne, G., 120

Caesar, L., 135, 145
Canale, M., 189
Carlson, L., 45
Cavar, D., 153, 173
Charniak, E., 29
Chiarcos, C., ix–x, xii, xv–xvi, 11–14, 34, 40, 48, 51, 60, 62, 117, 121, 127, 134, 152, 172–173
Chomsky, N., 27
Church, K., 28
Cimiano, P., 13–14, 53
Clahsen, H., 134
Clark, A., 47
Clear, J., 28
Cochran, P., 136
Cole, T., 112
Collins, M., 29
Cysouw, M., 41

Davison, S., 121
De Houwer, A., 187
de Melo, G., 13, 53
DeRose, S., 28
Dimitriadis, A., 62
Donohue, M., 50
Dryer, M., 41
Duerst, M., 4
Dunn, L. M., 190

Ďurčo, M., 104, 110, 111
Dye, C. D., 154, 158, 165

Eisenberg, S., 135, 138, 140
Erjavec, T., 28
Espinosa, L. M., 190
Esquinca, A., 189
Evans, J., 136
Evans, N., 41

Farrar, S. O., 19–22, 153
Ferrucci, D., 32
Finestack, L., 136
Fishman, J. A., 186–187
Flege, J. E., 187
Flynn, S., 155, 158–159, 164, 177, 193
Foley, C., 151, 158–160, 164, 193
Forkel, R., 57–58
Freudenthal, D., 134
Furbee, N. L., xv, 177

Gambino, C. P., 185
García, O., 190
Garside, R., 27
Gathercole, V. M., 190
Genesee, F., 187
Gómez-Pérez, A., xv, 13
Good, J., xv, 21, 43, 62
Gorman, K., 135–136
Grenoble, L. A., xv, 177
Grishman, R., 30
Grosjean, F., 187–188, 190
Gutierrez-Clellen, V., 189

Hamers, J. F., 187
Hammarstrom, H., 13, 41, 53
Haslhofer, B., 121
Hassanali, K., 136
Heidorn, P. B., 209
Heilmann, J., 136
Hellman, S., ix–x, 11, 40, 121, 152, 173
Herman, I., 14
Hey, T., 201
Hinrichs, E., 100
Hjelmslev, L., 22
Hux, K., 136

Ide, N., xvi, 28, 30–32, 34, 49, 51, 134, 173

Jones, D., 41

Kamholz, D., 44, 53, 62
Kang, C., 185, 191, 193
Kemp, K., 135
Kemps-Snijders, M., 34, 74
Kern, B., 202
King, G., 151, 154
King, T. H., 151–152
Kingsbury, P., 45
Kuĉera, H., 27, 44

Lambert, W. E., 187
Landes, S., 28
Langendoen, D. T., xvi, 20–21, 62, 77, 133–134, 153
Ledford, H., xiii–xiv
Lee, L., 136, 138
Lehmann, J., 7
Lewis, W. D., 19–22
Long, S., 136
Lubetich, S., 133, 148
Lupu, M., 28
Lust, B. C., ix, xiii–xv, 153–156, 158–159, 164–168, 173, 175–178, 186, 189, 191, 193

Mackey, W., 187
MacWhinney, B., xv, xvii, 131, 133, 152–153, 172, 174
Malvern, D., 144
Marcinkiewicz, M., xv, 27, 29
Marcus, G., 133, 149
Marcus, M., xv, 22, 27, 29, 38, 44
Matisoff, J., 50
Maxwell, M., 41–42
McCabe, A., 185
McCrae, J., 13–14, 53
McNew, G., 41
Meyers, A., 41, 45
Miller, E., 121
Miller, J. F., 136, 138, 147
Miller, K., 190
Moran, S., xvi, 41, 44, 48, 50, 52–53, 57, 172–173
Morgan, G., 194

Newman, R., 137, 144
Nivre, J., 32
Nordhoff, S., ix–x, 11–13, 40–41, 121, 152, 173

Otheguy, R., 190
Overton, S., 135–136
Owen, A., 145

Paradis, J., 187, 190
Pareja-Lora, A., ix, xii–xiii, xv, 13, 14, 60, 152–153, 156, 158, 172–173, 175
Pease, A., 20
Pederson, T., 61
Peña, E. D., 190
Pilar, D., 145
Pine J., 134
Pinker, S., 133
Pomerantz, J., 172–173
Poornima, S., 62
Powell, A., 118
Pradhan, S., 45
Prasad, R., 45
Price, L., 135–136
Prud'hommeaux, E., 49
Pustejovsky, J., 29, 45

Ratner. *See* Bernstein-Ratner, N.
Rehm, G., 42
Rice, M., 138–139
Rieger, O. Y., xvii, 172, 178, 201, 203, 205–206
Rispoli, M., 138
Rochon, E., 148

Sánchez, L., 186, 190
Santorini, B., xv, 27, 29, 44
Scarborough, H., 136, 138
Schalley, A., 61
Schuurman, I., 102
Shannon, C., 26
Silverman, S., 145
Simons, G., xiv–xv, xvii, 9, 20, 24, 101, 118, 132, 152–154, 173
Snow, C., xv, 131
Snyder, B., 52
Snyder, W., 176
Somashekar, S., 162, 164, 193
Southwick, S., 112
Sperberg-McQueen, M., 19
Squires, J., 189

Taylor, A., 44
Thomas, M., 189
Thordaardottir, E., 189
Tiedemann, J., 52
Tittel, S., 14
Trippel, T., xvi, 76, 107, 109–110, 132

Valian, V., 134
Verhagen, M., 32

Warburton, K., xvi, 83, 173
Weaver, W., 26
Wenger, E., 151, 176
Westerveld, M., 136, 147
Wexler, K., 134
Wilkinson, M., 3
Windhouwer, M., 51, 102, 104, 110–111, 113
Wong, A, 145
Wright, S., 50–51, 53, 71, 76, 173

Xia, F., 21

Yarowsky, D., 45

Zinn, C., xvi, 76, 105, 107, 109–110, 132
Zipser, H., 32
Zock, M., 44

Thematic Index

Accessibility, 2–3, 6–7, 10–11, 21, 42, 49–52, 57, 62, 100, 111. *See also* Addressability; Availability; Data: access [to]; Data, types of: accessible; FAIR principles; Findability; Interoperability; Repository: HTTP-accessible; Resolvability; Resource: accessibility; Reusability
Addressability, 107, 110. *See also* Accessibility; Findability; Resolvability
AI. *See* Artificial intelligence
Analysis, xi, xiii, xv–xvii, 6, 26, 29, 32–34, 43, 69, 77, 79–81, 87, 95, 131–137, 147, 152, 154–155, 161, 163–164, 171–173, 192, 194–196
  data, of (*see* Data)
  metadata, of (*see* Metadata)
  See also Cross-linguistic: analysis
Analysis, types of, 25, 47, 59, 134, 140, 143, 151, 158, 161, 171–173, 193
  linguistic (*see* Language/Linguistic)
  morphosyntactic, 25, 45, 52, 77, 133, 148 (*see also* Part of speech)
Annotation, xii, xv–xvi, 3, 8, 14, 21, 25–35, 43–44, 47, 51–52, 70, 101, 109, 131, 206
  graph, 31 (*see also* Format, types of: Annotation Graph)
  interoperability, 33, 35 (*see also* Interoperability)
  scheme, xiii, 20, 21, 26, 28, 31, 42, 73, 80, 88, 117 (*see also* Language/Linguistic: annotation)
  See also Data, types of: annotated; Gloss; Label; Label, types of; Linguistically annotated; Linguistic Annotation Framework; Metadata; Metatagging; Tag
Annotation, types of, 3, 30–31, 45–47, 80
  automatic, 27, 29, 153
  DTA, 158, 174–175 (*see also* Data Transcription and Analysis Tool)
  language/linguistic (*see* Language/Linguistic; Linguistically annotated)
  metadata (*see* Metadata)
  morphosyntactic, xv, 3, 27, 45–47, 80, 153 (*see also* Part of speech)
  named entity, 25, 35 (*see also* Named entity)
  ontological, xv, 44 (*see also* Ontology)
  part-of-speech (*see* Part of speech)
  semantic role, 25, 33, 35, 45 (*see also* Label, types of)
  sense, 25, 28 (*see also* Tag)
  standoff (*see* Standoff)
  syntactic, xv, 25, 27–28, 31, 42, 45–47
Aphasia, 132, 135, 148
AphasiaBank, 132, 135, 148
Archiving, 100–101, 113, 153, 172, 201–202, 208–209. *See also* arXiv; Data management, processes for: archiving [of]
Artificial intelligence (AI), x, 27, 59, 70
arXiv, xvii, 158, 203, 205–210. *See also* Archiving; Research data
Assessment, 43, 147, 192, 195, 205, 207
  clinical (*see* Clinical)
  of child language, 135–136, 138–140
Attribute, 3, 14, 30, 34, 79, 84, 88, 106, 120, 126. *See also* Feature
Authority file, 99, 107–108, 110, 113
  record, 107–109, 111

Authority file (cont.)
  See also Information, types of: authority file; Integrated Authority File; Virtual International Authority File
Automaticity, xvii, 25, 27, 29, 33, 35, 44, 47–48, 57, 59, 83, 131, 133, 135, 139, 142, 148, 153, 174–176, 192, 196
Availability, xiv, 1, 3, 7, 9, 14, 25, 27, 32, 41–44, 47, 48–49, 51–53, 57, 61, 71, 74, 79, 88, 90, 94–95, 101–102, 109, 111–114, 131–132, 134, 138–139, 147, 152–153, 158, 161, 165, 168, 172–174, 187, 192–193, 196, 206, 209. *See also* Accessibility

Bantu languages, 52
Basque, 52
Best practice, xiv, xvi, 2, 4, 10, 13, 29, 32, 48, 73, 92, 106, 112, 120, 122, 127–128, 151, 153, 155, 174, 202, 207. *See also* Practice; Recommendation
BIBFRAME, 112, 120. *See also* Library: resource; Library: standard; Resource: description
Big data, x, 39, 50, 101. *See also* Data; Data, types of
Bilingualism. *See* Multilingualism/ Multilinguality
British National Corpus, 28, 132
Brown Corpus, 27, 29, 44

Calibration, xii, xvii, 152–153, 155, 161–165, 168, 174–175, 192–193, 195–196
Cantonese, 76, 133, 142
Cape Verde Creole, 119
Catalan, 133
Category, x, 26, 28–29, 31, 33–35, 101
  data (*see* Data category)
  linguistic data (*see* Data category, types of)
  See also Descriptor; Markup; Metadata
CDB. *See* Concept, database
CES. *See* Corpus Encoding Standard
CHAT, 131–133, 135, 174–176. *See also under* Format, types of
Child, 132, 135–139, 142–144, 145–148, 153, 159–166, 168–169, 171–172, 174–175, 177, 186, 189–191, 193–195
CHILDES. *See* Child Language Data Exchange System
Child language, xv, 135–136, 138, 153, 165, 168, 172, 189
  ability, 135, 146, 191 (*see also* Child language: skill)
  data, 135, 153, 174–175 (*see also* Data; Language acquisition; Language/linguistic development)
  performance, 139, 142, 147, 189
  production, 159, 163, 165
  sample, 135, 148, 168
  skill, 135, 137 (*see also* Child language: ability)
  See also Child Language Data Exchange System; Language, types of: child
Child Language Data Exchange System (CHILDES), xv, 131–134, 136, 142, 144, 147–148, 152–153, 155, 172, 174–176, 196
Chinese, 28, 41, 133, 193
CLAN. *See* Computerized Language Analysis
CLARIN. *See* Common Language Resources and Technology Infrastructure
CLAVAS, 109, 112–113. *See also* Common Language Resources and Technology Infrastructure: vocabulary; Web: Service
Clinical, 135–136, 138–139, 146–148
  assessment, 135, 137, 145–147
  practice, 131, 136, 138 (*see also* Practice)
  See also Tool, types of: clinical
Clinician, 135–136, 142, 146–147, 177
CLLD. *See* Cross-Linguistic Linked Data
CMDI. *See* Component Metadata Infrastructure
CMDI2DC, 105. *See also* Component Metadata Infrastructure; Format, types of: CMDI; Format, types of: DCMI; Metadata: conversion; Web: service
Code, 1, 9, 48, 74, 88–90, 109, 118–120, 122, 125–126, 132–133, 162–163, 167, 196, 202, 209. *See also* Coding; Encoding
Code-switching, 132, 188, 191, 193–195
Coding, 134, 139, 153–154, 156, 161–167, 169, 171, 173, 175–176, 192–193, 195–196. *See also* Code; Encoding; Format; Markup
Cognitive science, 61, 151–152, 177, 186
Collaboration, xi, xiii–xiv, xvi–xvii, 6, 12, 21, 39–40, 52, 59, 61, 74, 95, 151–154, 158, 161, 172, 177, 196, 201–202, 208–210

COMEDI, 104. *See also* Tool, types of: CMDI metadata editor; Tool, types of: web-based

Comma-separated values (CSV), 4, 53, 57. *See also* Format; Format, types of

Common Language Resources and Technology Infrastructure (CLARIN), xvi, 34, 76, 78, 88, 99–107, 109–114, 133–134
  community, 104–107, 109–113 (*see also* Community)
  Component Registry, 104, 106–107, 109, 111–114 (*see also* Registry)
  Concept Registry, 34, 78, 102, 104–107, 110, 112–114 (*see also* Concept: registry)
  vocabulary, 106, 109 (*see also* CLAVAS; Vocabulary)

Community, xi–xiii, xvi, 2, 9, 10–14, 19–22, 34–35, 39, 42, 47, 51–52, 57, 59, 61, 73, 77–79, 82, 95, 99–101, 104–107, 109–113, 117–118, 120, 122, 127–128, 131, 153, 155, 158, 161, 173–174, 176, 185, 191, 193, 195, 202, 205–210
  of practice, 21, 77, 91 (*see also* Practice)
  See also Common Language Resources and Technology Infrastructure: community; Component Metadata Infrastructure: community; Language/Linguistic: community; Language/linguistic resource: community; Linguistics: community; Metadata: community; Ontology-Lexica Community Group; Open Language Archives Community; Research, types of: community; Resource, types of: community-supported; Semantic Web

Comparison, xvi–xvii, 9, 20, 31, 50–51, 61, 133, 135, 137, 139, 148, 151–153, 155, 158–165, 172, 177, 186–187, 189, 192. *See also* Cross-linguistic: comparability/comparison; Data: comparability [of]; Data processing, types of: comparison; Data, types of: comparable; Format: comparable

Component, 99, 102, 104, 109, 111–113
  CMDI (*see* Component Metadata Infrastructure: component)
  definition, 104–105, 108, 111, 113
  management, 102, 104, 106–107
  metadata, 100, 102, 105 (*see also* Component Metadata Infrastructure)
  registry (*see* Common Language Resources and Technology Infrastructure: Component Registry)
  See also Component Description Language

Component Description Language, 104

Component Metadata Infrastructure (CMDI), xvi–xvii, 99–114, 132
  -based metadata, xvi, xvii, 99–100, 104, 106–107, 109–111, 113–114 (*see also* Metadata)
  community, 106–107, 110–113 (*see also* Community)
  component, 102, 106–109, 111 (*see also* Component)
  instance, 104–107, 111
  interoperability, 99–100, 110, 113 (*see also* Component Metadata Infrastructure: semantic interoperability; Component Metadata Infrastructure: syntactic interoperability; Interoperability)
  metadata, 104, 107, 109–110 (*see also* Metadata; Metadata, types of)
  profile, 104–105, 107–108, 110–111 (*see also* Profile)
  semantic interoperability, 99, 111, 113 (*see also* Interoperability, types of: semantic)
  syntactic interoperability, 113 (*see also* Interoperability, types of: syntactic)
  vocabulary, 99, 107, 110–111 (*see also* Vocabulary)
  See also CMDI2DC; Infrastructure, types of: CMD/CMDI

Computerized Language Analysis (CLAN), 131–133, 136–137, 142, 143, 147, 174–176. *See also* MEGRASP; MOR; POST

Computer science. *See* Science, types of: computer

Concept, xiii–xiv, 6, 20–21, 43–44, 50–51, 53, 57, 70–71, 77–78, 81–82, 87–89, 91, 94, 104, 106, 107, 110–111
  Database (CDB), 73–74
  equivalent, 50, 110
  GOLD, 20–21 (*see also* GOLD)
  linguistic (*see* Language/Linguistic)
  registry, 99, 102, 106, 110, 113 (*see also* Common Language Resources and

Concept (cont.)
  Technology Infrastructure: Concept Registry; Registry)
  See also Data category: concept
Conceptual, xi, 6, 21–22, 26, 29, 48, 112, 138, 158, 173
  domain, 74–82, 83, 90, 92
  interoperability, 48–53 (*see also* Interoperability)
Conference on Natural Language Learning (CoNLL), 32, 131
CoNLL. *See* Conference on Natural Language Learning
Content, xi–xii, xvii, 4, 11, 31–34, 40, 44, 57, 69, 71–73, 76, 78–79, 83, 84, 86–87, 89, 94, 106, 120, 122, 126–127, 166, 173–174, 193, 202, 206–207, 209
  model, 72, 75
Context, xiv, 14, 134, 165–166, 168–169, 187, 189, 191
Copyright, 2, 42, 101, 207. *See also* Intellectual property; License; Right
Corpus, xv, xvii, 12–13, 21, 25–30, 33, 39, 42–47, 50–51, 57, 60–62, 69–70, 72–73, 77, 100–102, 104, 108–109, 111, 118, 121, 127, 131–135, 142–143, 148, 165–166, 168, 172–173, 177
  linguistics, xv, 25, 44, 177 (*see also* Linguistics, subfields of)
  tool (*see* SketchEngine)
  See also British National Corpus; Brown Corpus; Corpus Encoding Standard; Corpus, types of; TalkBank
Corpus Encoding Standard (CES), 30–31. *See also* Corpus; Encoding; Standard
Corpus, types of, 46, 51, 100–102
  annotated, 25–28, 39, 43, 45, 62 (*see also* Linguistically annotated: corpus)
  linguistic (*see* Language/Linguistic: corpus; Linguistically annotated: corpus)
  multimedia, 47 (*see also* Data, types of: multimedia/multimodal)
  parallel, 43, 45–47
  speech, 26, 43, 165–166, 168
  text, 43, 47, 101–102, 104
Cross-linguistic, 21, 40–41, 151, 156, 159, 161, 163–165, 173, 193
  analysis, xii, 158, 193 (*see also* Analysis)
  comparability/comparison, xvii, 152, 159–162, 164–167, 172 (*see also* Comparison)
  data, 62, 155, 174 (*see also* Data; Data, types of)
  difference, 159–160, 162, 192
  investigation/research/study, xii, 151–153, 155, 193, 196
  linked data (*see* Cross-linguistic Linked Data)
  search, 21, 62, 163
  similarity, 159, 162
Cross-linguistic Linked Data (CLLD), 50, 57–58, 61
CSV. *See* Comma-separated values
Czech, 28

Data, xi, 19, 48, 57, 131, 133, 142, 156, 159–160, 171–174, 176, 185, 187, 189, 201, 210
  access [to], 42, 49–50, 52, 99–100, 112, 40, 50, 113, 152, 154, 156, 165, 203, 211 (*see also* Accessibility; Data, types of: accessible)
  category (*see* Data category; Data category, types of)
  collection, 1–2, 13, 49–50, 61, 93, 132, 155–156, 190, 201–202 (*see also* Data management, processes for: collection; Dataset)
  comparability [of], 174 (*see also* Comparison; Data processing, types of: comparison; Data, types of: comparable; Format: comparable)
  complexity, 151, 154, 176, 193
  descriptor, 99, 102, 106–107, 110, 113 (*see also* Descriptor)
  entry, 155, 161, 176
  format, 3–4, 20, 33, 50, 53, 57, 111–112 (*see also* Format; Format, types of)
  interchange, 20, 77, 90 (*see also* Format, types of: interchange)
  interoperability [of], xi, 29, 39–40, 42, 50, 52, 56, 152, 173 (*see also* Data, types of: interoperable; Interoperability)
  label, 151, 161 (*see also* Label)
  management (*see* Data management; Data management, processes for)

ownership of, xiv, 207 (*see also* Intellectual property; License)
processing (*see* Data processing, types of)
provenance [of], 3, 7, 9, 11, 154, 156, 174, 176
provider [of], ix, 2, 10, 14, 108–109, 111–113 (*see also* Data: source)
quality, 151 (*see also* Data management, processes for: curation)
representation of, xi, 117, 154 (*see also* Encoding; Format; Format, types of)
source, x, 39, 47, 50, 52–53, 62, 109, 153, 195 (*see also* Data: provider)
structure, 43–44, 47, 51 (*see also* Data, types of: structured)
See also Data management; Data management, processes for; Data processing, types of; Data, types of
Databank, 19, 72, 132, 155, 174. *See also* Database
Database, ix, xii–xv, 5–9, 20, 47, 49–51, 56–57, 62, 69, 73–74, 78, 89, 99, 109–110, 118, 124, 132–135, 148, 152, 155, 165, 172–176, 195–196. *See also* Concept Database; Databank
Data category (DC), 34, 71–78, 80–82, 87–89, 90–94, 101–102, 106, 110, 119–120, 125, 173
concept, 82, 87 (*see also* Concept)
interchange format (*see* Data Category Interchange Format)
registry (*see* Data Category Registry)
repository (*see* Data Category Repository)
selection (DCS), 71–74
specification, xvi, 71–78, 85–89, 91–94
See also Data category, types of; Descriptor: elementary data; Standard: data category
Data Category Interchange Format (DCIF), 74, 79, 82, 83, 85–86. *See also* Data category; Format; Format, types of: interchange; MARTIF
Data Category Registry (DCR), xv–xvi, 51, 69, 73–74, 76, 79, 90, 92–94. *See also* Data Category Repository
Data Category Repository (DCR), xvi, 69, 75, 77–78, 80–82, 84, 88–90, 92–93, 95. *See also* Data Category Registry
Data category, types of, 75–76, 83, 93
CMDI (*see* Component Metadata Infrastructure)
complex, 74–76
experimental, 2, 158, 161–162, 164–165, 173–174, 195–196
linguistic, 70, 78, 88
primary, 2, 31–32, 44, 47, 192
private, 79, 86, 89
public, 79–80, 86
simple, 75, 83, 85, 90, 92
Data management, ix, xiii–xiv, xvi–xvii, 2, 14, 59, 105, 100, 111, 113, 151, 154–156, 174, 196, 201–202, 209, 211
import/export challenges, 175–176
standard, 2 (*see also* Standard)
See also Data; Data, types of; Data management, processes for; Data processing, types of
Data management, processes for, 59, 100, 111, 113, 151, 154, 156, 174, 201–202, 207
archiving [of], 113, 153, 159, 172, 177, 201–202, 209 (*see also* Archiving)
capture [of], 154, 158, 161, 173, 176 (*see also* Data management, processes for: collection; Data management, processes for: extraction; Data management, processes for: mining)
collection, xv, 155–156, 173, 201–202 (*see also* Data: collection; Data management, processes for: capture [of]; Data management, processes for: extraction; mining)
conversion, 34, 47, 56, 62, 83, 85–86, 111, 117, 126, 131, 134, 173–174
creation, xi, xiv, 25, 31, 40, 51, 59, 78, 153, 196
curation, xvi, 99, 106–107, 110–113, 209
documentation, 2, 3, 10, 153, 177, 191 (*see also* Documentation)
export, 132, 152, 172, 174–176
extraction, 151, 154, 201 (*see also* capture [of]; collection; mining)
integration, 21, 39, 47, 50, 62 (*see also* Integration)
(inter)linkage of, ix, xi, 3, 9, 12, 26, 121, 134, 152, 156, 161, 163, 173, 176, 195–196, 201, 209, 211 (*see also* Linguistic Linked Open Data; Linked Open Data: interlinkage)

Data management, processes for (cont.)
  mining, 77, 148, 152, 201 (*see also* Data management, processes for: capture [of]; Data management, processes for: collection; Data management, processes for: extraction)
  organization, 31, 156, 201–201
  preservation, xiv, 10–11, 154, 202, 209 (*see also* Preservation)
  reuse [of], xv, 15, 2, 61, 196, 201, 209, 211 (*see also* Reusability)
  sharing [of], ix, xii–xiv, xv–xvi, 20, 99, 111, 134, 152, 154–155, 173, 177, 190–191, 195–196, 201–202, 209 (*see also* Data, types of: sharable/shared; Share/sharing)
  storage, xiv, 111, 151, 156, 190
  transformation, ix, xviii, 42, 53, 58, 127, 172–173, 176
  use, xiii, 134, 142, 174, 201
  See also Data; Data management; Data processing, types of; Data, types of
Data processing, types of, xvi, 105, 100, 111, 132, 155, 161, 172–173, 175–176, 202
  analysis, 21, 151, 153–156, 159, 163–165, 171, 201–202 (*see also* Analysis)
  authentication, 135, 156, 201, 209
  comparison, 20–21, 152–153, 155, 159 (*see also* Comparison; Data: comparability [of])
  conversion, xvi, 3, 3, 34, 47, 56, 62, 81, 83, 86, 94, 111, 131, 134–135, 173 (*see also* Data management, processes for: transformation)
  dissemination, xi, xiv, xvi, 2, 40, 153–154, 172, 174, 201 (*see also* Data processing, types of: publication/publishing)
  publication/publishing, 50, 59–61 (*see also* Data processing, types of: dissemination)
  reanalysis, 154, 156, 159
  transformation, 56, 58, 172 (*see also* Data management, processes for: conversion)
  See also Data; Data, types of; Data management; Data management, processes for
Dataset, x, 2, 3, 5–9, 11, 20, 26, 39–40, 44, 47–48, 50, 56–59, 100, 105, 111–112, 119–120, 125, 127, 137, 152–153, 155–156, 161, 163, 168–169, 171, 175, 176–177, 195
  Semantic Web (*see* Semantic Web: dataset)
  See also Data collection
Data Transcription and Analysis Tool (DTA tool), xv–xvii, 132, 151–152, 155–156, 158–159, 163–165, 167–168, 171–177, 192, 194, 196
Data, types of, 3, 20, 30, 44, 47, 53, 58, 61–62, 79, 100, 112, 131, 161, 163–164, 169, 173–174, 176, 188–189, 194, 196
  accessible, 42, 52, 185 (*see also* Accessibility; Data: access [to])
  annotated, 29, 31, 43–44 (*see also* Annotation)
  big (*see* Big data)
  CMDI (*see* Component Metadata Infrastructure)
  CHAT (*see* CHAT)
  closed, 3 (*see also* Data, types of: private)
  comparable, xvi, 133, 163, 185, 196 (*see also* Comparison; Data: comparability [of]; Data processing, types of: comparison; Format: comparable)
  digital, xvi, 21, 42, 52, 152–153, 209
  experimental, 158, 164–165, 174
  freely available, 61 (*see also* Open)
  heterogeneous, 2, 9, 47
  -intensive research (*see* Research, types of: data-intensive)
  interoperable, xi, xiv, 39, 50, 56, 121, 173 (*see also* Data: interoperability [of]; Interoperability)
  language/linguistic (*see* Language/linguistic data)
  language acquisition (*see* Language acquisition data)
  lexical, 44, 53, 56, 58
  linguistically annotated (*see* Linguistically annotated)
  linked (*see* Linked Data)
  linked open (*see* Linked Open Data)
  multilingual (*see* Multilingual (adjective): data/information)
  multimedia/multimodal, 154, 194, 196 (*see also* Corpus, types of: multimedia)
  open (*see* Data, types of: freely available; Open Data)
  private, 3 (*see also* Data, types of: closed; Private data)
  RDF (*see* Resource Description Framework)
  research (*see* Research data)

  sharable/shared, ix, 20, 152, 155, 173, 177 (*see also* Data management, processes for: sharing [of]; Share/sharing)
  structured, 1, 7, 57, 107, 163, 202
  TalkBank (*see* TalkBank)
  See also Data; Data management; Data management, processes for; Data processing, types of
DatCatInfo, xv–xvi, 69, 71, 78, 87, 92–94. *See also* Data Category Repository
DBPedia, 5–7, 50, 56, 59, 126–127
DC. *See* Data category
DCIF. *See* Data Category Interchange Format
DCMI. *See* Dublin Core Metadata Initiative
DCR. *See* Data Category Registry; Data Category Repository
DCS. *See* Data category: selection
Description Logics, 50. *See also* Logic; OWL-DL
Descriptor, 99, 102, 104–106, 108
  elementary data, 99, 102
  metadata, 100, 102, 105–106
  values, 102, 107–108, 113
Developmental Sentence Score (DSS), 133, 136–144, 148
Dictionary, 19–20, 39, 41–44, 47, 53, 56, 62. *See also* Lexicon
Digital Object Identifier (DOI), 132–133
Digraph. *See* Graph, types of: directed
Discourse, 25, 118, 122–123, 134, 167, 189, 193
Discoverability, 4, 6, 10, 59, 107, 117–118, 121, 126, 151, 161, 164, 201, 209, 211. *See also* Findability
DOBES. *See* Dokumentation Bedrohter Sprachen
Document Type Definition (DTD), 30
Documentation, 2, 10–11
  data (*see* Data management, processes for: documentation)
  language (*see* Language/Linguistic: documentation)
  metadata (*see* Metadata: documentation [of])
  vocabulary (*see* Vocabulary)
  See also DOBES
Documentation of Endangered Languages. *See* Dokumentation Bedrohter Sprachen
DOI. *See* Digital Object Identifier
Dokumentation Bedrohter Sprachen (DOBES), xiv–xv, 20
DSS. *See* Developmental Sentence Score
DTA tool. *See* Data Transcription and Analysis Tool
DTD. *See* Document Type Definition
Dublin Core Metadata Initiative (DCMI), 101, 104–106, 109–113, 118, 120, 122. *See also* CMDI2DC; DCMI2DC; Format, types of: bibliographic; Format, types of: DCMI; Library: standard; Metadata standard: Dublin Core
Duplication, 70, 73, 77, 79, 86–88, 90, 94, 106, 113. *See also* Redundancy
Dutch, 44, 133, 142
  Pennsylvania Dutch (*see* German)

EAGLES, 30, 34
Electronic Metastructure for Endangered Languages (E-MELD), xv, 19–21, 62, 134, 153
E-MELD. *See* Electronic Metastructure for Endangered Languages
Encoding, xvi, 4, 20, 22, 30, 39, 42, 50, 57, 101, 176
  scheme, 23, 118, 120, 122 (*see also under* Annotation)
  See also Coding; Corpus Encoding Standard; Data Transcription and Analysis Tool; Text Encoding Initiative; Transcription
English, 6, 27, 41, 43, 76–77, 84, 93, 126, 133–135, 140, 142, 145, 148, 158–160, 162–169, 171–172, 174, 189–193
  dialect/variant of, 27, 143, 190–191
E-science, 201–203, 205, 209, 211. *See also* Open: science
Etruscan, 52
EVAL, 135, 148. *See also* KIDEVAL
eXtensible Markup Language (XML), xii, 4–5, 10, 14, 19–20, 30–31, 49, 57, 71, 84, 85–88, 104, 106, 111, 117, 127, 133
  format, 47, 118, 134
  markup, 72, 79, 120
  schema (*see* XML Schema)
  See also Format, types of: XML; Language, types of: XML; Markup: XML; Schema: XML; Standard: XML; Vocabulary, types of: XML; XML Schema

FAIR principles, 3, 9–10. *See also* Accessibility; Findability; Interoperability; Reusability
Faroese, 52
Feature, xvi, 19–20, 25–26, 28–32, 57–59, 126
  language/linguistic (*see* Language/Linguistic)
  structure (FS), 19–20
  system declaration (FSD), 19
Federated Content Search, 99. *See also* Federation
Federation, 49–51. *See also* Federated Content Search
Findability, 3, 10. *See also* Accessibility; Addressability; Discoverability; FAIR principles; Interoperability; Resolvability; Reusability
Fluency, 132, 147, 191, 194–195
Form, 22, 43, 107. *See also* Word
Format, xii–xiv, xvi, 3–4, 6–7, 9–11, 21, 28–30, 32–33, 42, 48–50, 53, 75, 101, 111, 117–119, 125, 128, 131, 135, 154, 174, 176
  comparable, 48–49 (*see also* Comparison; Data: comparability [of]; Data processing, types of: comparison; Data, types of: comparable)
  conversion [of], 34, 47, 56, 62, 83, 85–86, 110–111, 117, 126, 131, 134, 173
  See also Coding; Encoding
Format, types of, 3, 7, 10, 20–21, 29–32, 42, 48–50, 53, 56, 72, 110–111, 119, 125, 132, 211
  AG (*see* Format, types of: Annotation Graph)
  Annotation Graph (AG), 31 (*see also* Format, types of: graph)
  bibliographic, 111 (*see also* Format, types of: DCMI; Format, types of: MARC 21)
  CHAT, 131–132, 135, 174–175
  CMDI (*see* Component Metadata Infrastructure)
  CoNLL (*see* Conference on Natural Language Learning)
  CSV (*see* Comma-separated values)
  data (*see* Data: format)
  DCMI, 105, 110–111 (*see also* Dublin Core Metadata Initiative; Format, types of: bibliographic)
  digital/electronic, 42, 73, 75
  graph, 48 (*see also* Format, types of: Annotation Graph)
  IGT (*see* Interlinear glossed text)
  in-line, 30–31
  interchange, 10, 14, 20, 32, 71, 74, 79, 117 (*see also* Data Category Interchange Format; MARTIF)
  JSON-LD (*see* JavaScript Object Notation: for Linked Data)
  machine-readable, 10, 43, 71 (*see also* MARC 21; MARTIF)
  MARC 21, 110–111 (*see also* MARC 21)
  MARTIF (*see* MAchine-Readable Terminology Interchange Format)
  metadata (*see* Metadata: format)
  OLAC, 101 (*see also* Open Language Archives Community)
  open (*see* Open: format)
  output, 53, 132
  OWL, 35 (*see also* Web Ontology Language)
  PDF (*see* Portable Document Format)
  Penn Treebank, 29 (*see also* CoNLL; Penn Treebank)
  physical, 26, 28–29, 31–32, 34
  RDF (*see* Resource Description Framework)
  standard (*see* Standard: format)
  standardized, 4, 154 (*see also* Standard)
  standoff (*see* Standoff)
  syntactic (*see* Syntactic)
  TBX (*see* TermBase eXchange)
  transcription (*see* Transcription)
  XML, 47 (*see also* eXtensible Markup Language)
Fourth Paradigm, 201–202
French, 28, 41, 79, 133, 142, 148, 158, 160, 162–164, 193
FS. *See* Feature: structure
FSD. *See* Feature: system declaration

Gemeinsame Normdatei (GND). *See* Integrated Authority File
General Ontology for Linguistic Description (GOLD), xv–xvi, 19–22, 62, 77, 88–89, 126, 133, 153–154. *See also* Concept: GOLD; OntoLingAnnot; Ontologies of Linguistic Annotation; Ontology: linguistic; OntoTag

German, 6, 28, 32, 41, 44, 76, 79, 109, 122, 133–134, 142
  Pennsylvania Dutch, 44
GitHub, 132, 209
Gloss, 3, 19, 39, 43–44, 47, 53, 56, 62, 166. *See also* Annotation; Tag
Glottolog, 41, 43, 53
GND. *See* Gemeinsame Normdatei; Integrated Authority File
GOLD. *See* General Ontology for Linguistic Description
Grammar, 19–20, 41, 43, 45, 100–101, 133, 168–169, 189. *See also* Parse; Parser
Graph, 5, 31–32, 44, 48, 51, 53, 56–57, 62, 121, 173. *See also* Graph, elements; Graph, types of
Graph, elements
  arc, 121 (*see also* Graph, elements: edge)
  directed edge (*see* Graph, elements: arc)
  edge, 31, 48 (*see also* Graph, elements: arc; Graph, elements: directed edge)
  node, 5, 20, 48, 57, 108, 121
Graph, types of
  annotation (*see* Annotation: graph)
  directed, 5, 32, 48, 121 (*see also* Digraph; Graph: labeled directed)
  labeled directed, 48, 51 (*see also* Graph, types of: directed)
  RDF (*see* Resource Description Framework: graph)
  translation (*see* Translation: graph)
Gujarati, 190

Haitian Creole, 190, 193
Handle, 102, 132. *See also* Permanent identifier; Persistent identifier
Harmonization, x, xvi, 6, 34, 62, 71, 73, 82, 87–88, 90, 92, 94–95, 101, 111, 133. *See also* Data processing, types of; Language/linguistic resource; Metadata; Standardization
Hausa, 42
Hebrew, 52, 133
HTML. *See* Hypertext Markup Language
HTTP. *See* Hypertext Transfer Protocol
Humanities, xvii, 1, 3, 61, 99–100, 151
Hypertext Markup Language (HTML), 5, 15, 51, 127, 132–133
Hypertext Transfer Protocol (HTTP), 4, 48–49, 51, 107, 121, 124–127

Icelandic, 52
Identity management, 108. *See also* Authority file; IRI; ISNI; ORCID; PID; ResearcherID
IGT. *See* Interlinear glossed text
Index of Productive Syntax (IPSYN), 133, 136–140, 142–144, 148
Individual (person), xvii, 11, 33, 61, 117–118, 152, 172, 186–187, 192, 194. *See also* Subject (participant in study)
Indonesian, 133
Inflection, 165, 168–169, 172
Information, x, 4–6, 8, 20, 25–26, 29–34, 41, 44, 47, 48–51, 53, 58–59, 69–71, 73, 76–78, 80, 83, 86, 88, 90, 92, 99, 102, 107–111, 117–121, 125, 132–133, 154, 156, 161, 165, 174, 187–189, 191–193, 196, 201–202, 205–206, 209
  distribution of, x, 73, 155
  extraction of, 30, 134, 161
  integration [of], x, 48–51 (*see also* Federation)
  loss [of], 31, 79, 111
  science (*see* Science)
  source of, 88, 90, 121
  technology, 42, 52, 61, 201, 206
Information, types of, xiv, 6, 26–27, 29, 31, 33, 53, 70, 75–77, 84, 86, 110, 124, 131, 133, 154–156, 159, 166, 177, 188–189, 193–194
  authority file, 99, 107–108, 110 (*see also* Authority file)
  language/linguistic (*see* Language/Linguistic: information)
Infrastructure, x–xi, xiv, 7, 11, 14, 19, 47, 51, 99–101, 118, 128, 151, 172, 177, 202–203
Infrastructure, types of, xi–xiii, xvii, 59, 203, 205
  centralized, 100–101, 111
  CLARIN (*see* Common Language Resources and Technology Infrastructure)
  CMD/CMDI, 100, 104–105, 107, 111–112, 114 (*see also* Component Metadata Infrastructure)

  cyberinfrastructure, ix, xi, 151, 154–155, 158, 196
  digital, xvi, 21, 172–173
  distributed, 100–101, 111
  e-science (*see* E-science)
  information (*see* Information)
  Linked Data (*see* Linked Data)
  metadata (*see* Component Metadata Infrastructure)
  OLAC (*see* Open Language Archives Community)
  research (*see* Research)
  research data (*see* Research data)
  shared, 203 (*see also* Share/sharing)
  sustainable, 210 (*see also* Sustainability)
  technical, 8, 208
Integrated Authority File (GND), 108–109. *See also* Authority file; Gemeinsame Normdatei
Integration, 56. *See also* Data management, processes for: integration; Information: integration [of]; Resource: integration
Intellectual property, xiii, 209. *See also* Copyright; Data: ownership of; License; Right
Interdisciplinarity, xiv, 51–52, 59, 61, 151, 155, 172, 176–177
Interlinear glossed text (IGT), 19–20, 44, 47
Internationalized Resource Identifier (IRI), 4. *See also* Identity management
International Organization for Standardization (ISO), 30–31, 44, 65, 69, 70–74, 76–78, 82, 88–92, 94–95, 99–102, 106, 108–109, 111–113, 118, 120, 122, 126, 215
  Central Secretariat, 73–74, 92
  Online Browsing Platform (OBP), 74, 88, 90
  ISOcat, xv–xvi, 34, 51, 69, 74, 76, 79, 92, 94, 102, 106 (*see also* Data Category Registry; DatCatInfo)
  Registration Authority, 74, 95 (*see also* Registration Authority)
  Technical Committee 37 (ISO/TC 37), xii, xv, 69–71, 73–74, 76, 90, 94–95, 173 (*see also* International Organization for Standardization: Technical Committee 37, Subcommittee 4)
  Technical Committee 37, Subcommittee 4 (ISO/TC 37/SC 4), 34, 39, 51, 73 (*see also* International Organization for Standardization: Technical Committee 37)
International Standard Name Identifier (ISNI), 108. *See also* Identity management
Interoperability, x–xiii, xvi–xvii, 10, 14, 32–33, 35, 40, 42–43, 47, 48, 52, 56, 61, 73, 99, 110, 112, 117, 152, 177, 202, 208. *See also* Accessibility; FAIR principles; Findability; Interoperability, types of; Reusability
Interoperability, types of, 52
  annotation (*see* Annotation: interoperability)
  CMDI (*see* Component Metadata Infrastructure: interoperability)
  conceptual (*see* Conceptual: interoperability; Interoperability, types of: semantic)
  data (*see* Data: interoperability [of]; Data, types of: interoperable)
  language resource (*see* Language/linguistic resource: interoperability [of])
  semantic, xvi–xvii, 33–34, 50, 99, 105, 110–111, 113, 120 (*see also* Conceptual)
  structural, 48–53, 56
  syntactic, xvi, 33–35, 106
Investigation, xiii, 40, 194, 202. *See also* Research
  cross-linguistic (*see* Cross-linguistic: investigation/research/study)
IPSYN. *See* Index of Productive Syntax
IRI. *See* Internationalized Resource Identifier
ISNI. *See* International Standard Name Identifier
ISO. *See* International Organization for Standardization
ISOcat. *See under* International Organization for Standardization
ISO/TC 37. *See* International Organization for Standardization: Technical Committee 37; International Organization for Standardization: Technical Committee 37, Subcommittee 4
Italian, 28, 133

Japanese, 133
JavaScript Object Notation (JSON), 10, 32
  for Linked Data (JSON-LD), 4, 13, 32

JSON. *See* JavaScript Object Notation
JSON-LD. *See* JavaScript Object Notation: for Linked Data

KIDEVAL, xii, xvii, 135, 136–138, 142–144, 148. *See also* EVAL; Tool, types of: clinical
Klingon, 44
Knowledge, x–xi, xiii–xv, 1, 6, 14, 31, 41, 69, 71, 163–164, 173, 176, 201
  base, 20, 59
  of language, 151–152, 155, 188–189
  representation, x, xii, 3, 59, 214
Korean, 189

Label, 5–6, 31, 33, 41, 43, 50, 78, 89, 175–176, 196. *See also* Annotation
Label, types of, 173, 175
  annotation (*see* Annotation; Annotation, types of)
  data (*see* Data)
  metadata (*see* Metadata: label/labelling [of])
  named entity (*see* Annotation, types of: named entity)
  ontological, 173 (*see also* Annotation, types of: ontological; Ontology)
  RDF (*see* Resource Description Framework)
  semantic role, 25, 35 (*see also* Annotation, types of: semantic role; Semantic role)
LAF. *See* Linguistic Annotation Framework
Language. *See* Language/Linguistic
Language acquisition, xi–xii, xv, xvii, xvii, 2, 134, 148, 151–152, 154–155, 158–159, 161–162, 164, 176–177, 185, 187, 189, 190, 194
  research/study of, 156, 164, 176
  See also Language/linguistic development
Language acquisition, types of, 188, 190, 196
  child, 132, 153, 190, 194 (*see also* Child language development)
  data, 151, 159
  first (L1), 152, 157, 177 (*see also* Child language)
  multilingual (*see* Multilingual (adjective): language acquisition)
  second (L2), 132, 152, 177, 185, 189
Language Applications (LAPPS), 32, 34, 134
Language Archive, the (TLA), 2, 132, 134, 152, 192. *See also* Open Language Archives Community)
Language Environment Analysis (LENA), 131, 196
Language/Linguistic, xii–xiii, xvii, 3, 19, 25, 27–29, 31, 33–35, 43, 47, 61, 77, 80–82, 131–135, 138, 142, 147–148, 151–156, 158–159, 161–165, 171, 174, 176–177, 185–194, 196, 213
  acquisition of (*see* Language acquisition; Language/linguistic development)
  analysis, 20, 22, 50, 133, 153–154, 171 (*see also* Analysis)
  annotation, xii, xv, 14, 19–20, 22, 25, 26–28, 30–34, 44–45, 101, 151, 155, 173 (*see also* Annotation; Annotation: scheme; Linguistically annotated)
  community, 19, 73, 77, 196 (*see also* Community)
  competence, 187–189 (*see also* Language/Linguistic: proficiency; Language/linguistic skill, type of)
  concept, 78, 88, 94 (*see also* Concept)
  context, 165, 168–169 (*see also* Context)
  corpus, 61, 134–135 (*see also* Corpus; Corpus, types of)
  data (*see* Language/linguistic data)
  data category (*see* Data category, types of: linguistic)
  description, xv, 25, 57, 153 (*see also* Language/Linguistic: documentation)
  development (*see* Language acquisition; Language/linguistic development)
  diversity, 40–41, 50, 58
  documentation [of], xv, 40–43, 62, 101, 177, 190 (*see also* Language/Linguistic: description; Documentation)
  domain, 91, 94, 189
  feature, xvii, 58–59, 80, 131, 134, 138, 140 (*see also* Language/Linguistic: property)
  field, 118, 122–123
  information, 30, 34, 49, 122, 154, 161 (*see also* Information)
  impairment, 136, 148, 153, 177
  knowledge of (*see* Knowledge)
  Linked Data (*see* Linguistic Linked Data)
  Linked Open Data (*see* Linguistic Linked Open Data)

Language/Linguistic (cont.)
  metadata, 57, 58, 100, 109–110, 113, 121 (*see also* Metadata)
  markup, xi, xiv, 10, 19, 62, 151, 155, 172, 176–177 (*see also* Annotation; Markup)
  performance, 139, 142, 147, 188–189 (*see also* Language/linguistic skill, types of: production)
  processing, 185, 188, 194
  production [of] (*see* Language/linguistic skill, types of)
  proficiency, 186–189, 191 (*see also* Language/Linguistic: competence)
  profile, 186, 188, 191 (*see also* Language/linguistic resource: profile; Profile)
  property, 153–154, 158 (*see also* Language/Linguistic: feature)
  query, 48–50
  resource (*see* Language/linguistic resource; Language/linguistic resource, types of)
  resource description (*see* Language/linguistic resource: description [of])
  science (*see* Science)
  Section, 73–76, 79–82
  structure, 161, 164–165, 167, 187
  technology, 3, 26, 101
  test, 189
  theory, x, xiii, 25, 27–28, 152
  Type, 118, 122–123
  use of, 151, 185, 187–190 (*see also* Language/Linguistic: performance; Language skills, types of: production)
  utterance, 26–27, 161
  See also Language, types of
Language/linguistic data, x–xvii, 2, 7, 10, 13, 19, 21–22, 25–27, 30, 33–34, 39–40, 42–44, 48, 50, 52–53, 58–62, 69, 102, 110, 118, 121, 132, 134–136, 139, 152–155, 158–159, 161–162, 164–165, 172–174, 185, 188, 196
  processing (*see* Language/Linguistic)
  source, 39, 62, 153
Language/linguistic development, 135, 142, 144, 160, 163–165, 172, 185–186, 189, 195
Language/linguistic resource, x, xii–xiii, xvii, 10–11, 13, 25, 32, 34, 39–40, 42–44, 46–47, 51–52, 59–62, 69–70, 73–74, 77, 80, 83, 88, 90–91, 93–94, 99–102, 105, 107, 112–114, 117–118, 121, 125–126, 173–174
  annotated, 25, 32–33 (*see also* Annotation; Annotation, types of; Linguistically annotated)
  community, 39, 42 (*see also* Community)
  data category (*see* Data category, types of)
  description [of], 101, 117, 121, 125, 127–128
  harmonization, 62 (*see also* Harmonization)
  interoperability [of], x, 83 (*see also* Interoperability)
  management [of], 34, 73–74
  metadata, 14, 48, 117, 126 (*see also* Metadata)
  profile, 77, 102 (*see also* Profile)
  reuse [of], x, 40, 47 (*see also* Reusability)
  sharing [of], 118 (*see also* Share/sharing)
  Switchboard, 99
  See also Lexical resource; Resource
Language/linguistic sample, 136, 138, 140, 147–148, 168
Language/linguistic skill, type of
  comprehension, 187–189, 191
  expressive, 135, 188
  production, 159, 161, 163, 165, 189 (*see also* Language/Linguistic: performance; Language/Linguistic: use of)
  reading, 187–188
  speaking, 187–188, 195
  writing, 187–188
Language Sample Analysis (LSA), 135–140, 142, 145–148
Language, types of, xiii, 9, 30, 21, 41–44, 47, 72, 75–77, 106, 126, 132, 138–139, 142, 185, 188–190, 193
  adult, 160, 168–169
  African, 190
  Bantu, 52
  Caucasus, 62
  child, xv, 132, 135–138, 140, 146–148, 153, 163, 165, 168, 172, 174–175, 189 (*see also* Child language)
  Eastern, 28
  endangered, xv, 20, 41, 62, 101, 177 (*see also* Language, types of: under-resourced)
  European [Union], 28, 43
  first (L1), 152, 177, 189, 190, 192 (*see also under* Language acquisition, types of)

  human, 26–27, 151 (*see also* Language, types of: natural)
  Indo-European, 193
  less-resourced, 52, 61–62 (*see also* Language, types of: low(er)-density; Language, types of: under-resourced; Language, types of: weakly supported)
  low(er)-density, 41 (*see also* Language, types of: less-resourced; Language, types of: under-resourced; Language, types of: weakly supported)
  medium density, 41 (*see also* Language, types of: under-resourced)
  markup (*see* Markup)
  minority, 186, 188, 190 (*see also* Language, types of: under-resourced)
  natural, x, xii, xv, 10, 26, 47, 69, 134, 152, 158, 173 (*see also* Language, types of: human)
  query, 4, 9, 48–50, 128
  RDF-based (*see* Resource Description Framework)
  second (L2), 132, 135, 152, 185–189, 192 (*see also under* Language acquisition, types of)
  Semitic, 52
  sign, xiii, 194–196
  spoken, 56–57, 101, 131, 133–135, 137, 194
  Turkic, 62
  under-resourced, 39–43, 47, 48, 50, 52–53, 56–59, 61–62 (*see also* Language, types of: endangered; Language, types of: less-resourced; Language, types of: low(er)-density; Language, types of: medium density; Language, types of: minority; Language, types of: weakly supported)
  weakly supported, 42–43 (*see also* Language, types of: less-resourced; Language, types of: low(er)-density; Language, types of: under-resourced)
  Western, 28
  XML (*see* eXtensible Markup Language)
LAPPS. *See* Language Applications
LAPSyD. *See* Lyon-Albuquerque Phonological Systems Database
lemon. *See* Lexicon Model for Ontologies
LENA. *See* Language Environment Analysis
Lexical resource, 13, 39, 43–44, 53, 56, 59, 61–62, 100–101. *See also* Language/linguistic resource; Resource
Lexicon, 21, 44, 47, 52–53, 56, 60, 118, 122–123, 173, 188. *See also* Data, types of: lexical; Dictionary; lemon; Lexical resource; Ontology-Lexica Community Group
Lexicon Model for Ontologies (lemon), 53, 57
Library, 108–110, 112–113, 201–202
  catalog, 100, 109–110, 114
  resource, 112 (*see also* Resource: description)
  standard, 112 (*see also* BIBFRAME; DCMI; MARC 21)
  world, 101–102, 110, 112–113 (*see also* Context)
License, 1–3, 6–7, 11, 50, 59–61, 95, 100, 209, 211
  open (*see* Open: license)
  See also Copyright; Data: ownership of; Intellectual property; Open: resource; Open: source; Open Data; Right
Linguist, xii, 3, 14, 20, 22, 27, 39, 42, 50, 126, 177
Linguistically annotated, 53, 59
  corpus, 26–27
  data, 25, 30, 44
  resource, 25, 27, 33
Linguistic Annotation Framework (LAF), 31–32
Linguistic Data Consortium (LDC), 1, 134
Linguistic Linked Data (LLD), 39, 173. *See also* Linguistic Linked Open Data; Linked Data; Linked Open Data; Open Data; Resource, types of: Linguistic Linked Open Data; Resource, types of: Linguistic Open Data
Linguistic Linked Open Data (LLOD), x–xi, xvi–xvii, 1–2, 4–5, 7, 9–10, 12–14, 39–41, 51, 56, 59–63, 131, 134–135, 173, 185, 195–196, 210–211
  cloud, x, xii, 11, 40, 47, 52, 56, 59–61, 117, 121, 127, 214
  cloud diagram, 2, 12–13, 59, 61
  development of, x, xvi–xvii, 185, 191
  ecosystem, 39, 51, 63
  resource, ix, 12–13, 51
  vision, xii, xiv, 185–186, 195–196
  See also Linguistic Linked data; Linked Data; Linked Open Data; Open Data; Resource

Linguistics, xi–xii, xiv, xvi, 1–3, 6, 10, 12, 14, 20–22, 26, 40, 43, 47, 59–61, 73, 87, 90, 118, 126, 151–153
  community, 19–20, 61, 73 (*see also* Community)
  corpus (*see* Corpus linguistics)
Linguistics Society of America (LSA), ix, 40
Linguistics, subfields of, 45, 61, 117, 135
  applied, x–xii, xiv
  computational, xii, xv, 61
  corpus (*see* Corpus linguistics)
  lexicography, x, xv, 61
  literature, 87
  morphology, 45, 73, 131, 133, 154, 164
  morphosyntax, 21, 34, 45, 77, 80–82
  phonology, 57, 119, 126, 131
  syntax, 26–27, 33–34, 121–122, 125, 131, 133–135, 152, 174, 176
LinguistList, the, xv, 20, 127
Linked Data, xv–xvii, 4, 6–8, 11–14, 26, 31, 39–41, 47–52, 56–59, 62, 99–100, 107, 109–114, 117, 120–122, 126–128, 152, 173, 201–202
  cloud, 53, 56–57, 114
  framework, 40, 117, 120–121
  linguistics, in, x, 12, 39, 51, 59, 61
  paradigm, 52, 121, 128
  rule of, 117, 122, 124, 126–127
  technology, 39–40, 62
  See also Linguistic Linked Data; Linguistic Linked Open Data; Linked Open Data
Linked Open Data (LOD), x–xii, xvii, 1–2, 7, 9, 14, 40, 48, 59, 61, 99, 102, 108, 111, 152, 172–174, 185, 210–211
  cloud, x, xii, 7, 59, 173–174 (*see also* Linked Open Data: network)
  framework, xii, 172, 176–177, 179
  interlinkage, 172
  network, xiii–xiv, 49, 131 (*see also* Linked Open Data: cloud)
  paradigm, x, 48, 61
  technology, xv, 39, 62
  See also Linguistic Linked Open Data; Linked Data; Linked Open Data, types of; Open Data
Linked Open Data, types of
  multilingual, 12–13, 39 (*see also* Multilingual (adjective))
Linked Open Dictionaries (LiODi), 53, 62
LiODi. *See* Linked Open Dictionaries
LLOD. *See* Linguistic Linked Open Data
LOD. *See* Linked Open Data
Logic, 22. *See also* Description Logics; OWL-DL
  first-order, 22
LSA. *See* Language sample analysis; Linguistics Society of America
Lyon-Albuquerque Phonological Systems Database (LAPSyD), 118, 124, 126

MAchine-Readable Cataloging (MARC), 102, 110. *See also* Library: standard; Metadata standard
MAchine-Readable Terminology Interchange Format (MARTIF), 10. *See also* Data: interchange; Data Category Interchange Format; Format, types of: interchange
Maintainability, 111. *See also* Sustainability
Management
  data (*see* Data management; Data management, processes for)
  identity (*see* Identity management)
  metadata (*see* Metadata: management [of])
Mandarin, 142
Mapping, 4, 31–32, 34, 43, 57, 59, 78, 83, 85, 106, 110–111, 113, 123, 172, 175
MARC 21. *See also* MAchine-Readable Cataloging
Markup, xiii, xvii, 10, 19–20, 30, 71–72, 78–79, 83, 85, 117, 120, 154, 158–162, 164, 192–196
  language/linguistic (*see* Language/Linguistic)
  XML (*see* eXtensible Markup Language)
  See also Annotation; Coding; Label; Tag
MARTIF. *See* MAchine-Readable Terminology Interchange Format
Max Planck Institute for Psycholinguistics (MPI), 69, 74, 76, 78–79, 95
Mean length of utterance (MLU), 133, 136–140, 143, 146–148, 161, 165, 167–168, 177
MEGRASP, 133, 137. *See also* CLAN; MOR; Part of speech; POST; Tagger
Metadata, xv–xvii, 3, 7, 9–11, 13, 35, 44, 48, 52–53, 57, 59, 78, 87, 99, 101–102, 104–106,

109–112, 114, 119–120, 122, 127–128, 132, 135, 151, 153–156, 158–159, 161, 169, 171, 173–176, 186, 190–192, 195–196, 202, 209
  analysis [of], 155, 159 (*see also* Analysis)
  annotation [of], 155 (*see also* Annotation)
  collection [of], xvii, 11, 13, 61, 75, 155, 186, 190, 196
  community, 128 (*see also* Community)
  component (*see* Component: metadata)
  conversion, 105, 110–112 (*see also* CMDI2DC)
  description [of], 100–101, 105, 110
  descriptor (*see* Descriptor)
  documentation [of], 154, 186, 196 (*see also* Documentation)
  element, 106, 120–122
  format, 102, 113, 118
  harvest, 101, 105, 127
  label/labelling [of], 151, 161
  management [of], xvii, 105, 111, 155
  modeling [of], 99, 104, 106
  profile, 102–104 (*see also* Metadata: set; Profile)
  provider [of], 104–105, 109–110
  publication/publishing [of], 105, 134–135
  record, xvii, 112, 117, 118–119, 127
  repository, 52, 118, 124
  scheme, 99, 102, 106, 110–111, 113 (*see also* Scheme)
  set, 101–102, 104, 106 (*see also* Metadata: profile)
  standard (*see* Library: standard; Metadata standard)
  vocabulary, 107, 111–112 (*see also* Vocabulary, types of)
  See also Component Metadata Infrastructure; Dublin Core Metadata Initiative
Metadata standard, 99–100, 110
  Dublin Core, 110 (*see also* Dublin Core Metadata Initiative)
  MARC 21, 110 (*see also* MAchine-Readable Cataloging)
  OLAC, 117–118, 126, 128 (*see also* Open Language Archives Community)
  See also Library: standard
Metadata, types of, 107–109
  bibliographic, xvi, 99, 110, 112
  CMDI-based (*see* Component Metadata Infrastructure: -based metadata; Component Metadata Infrastructure: metadata)
  creator, 107, 110–111
  Dublin Core (DC), 101, 104 (*see also* Dublin Core Metadata Initiative)
  institution, 107, 109–110
  language/linguistic (*see* Language/Linguistic)
  METANET, 42–43
  participant, 192, 195 (*see also* Metadata, types of: person-related; Metadata, types of: researcher)
  person-related, 107–110 (*see also* Metadata, types of: participant; Metadata, types of: researcher)
  researcher, 108–110 (*see also* Metadata, types of: person-related; ORCID; ResearcherID)
  research, 100, 102, 108–109
  resource (*see* Resource: metadata)
Metatagging, 20. *See also* Scheme; Tag
MLODE. *See* Multilingual Linked Open Data for Enterprises
MLU. *See* Mean length of utterance
MOR, 133, 137. *See also* CLAN; MEGRASP; Part of speech; POST; Tagger
Morpheme, 3, 21, 141, 167–168, 180, 193
MPI. *See* Max Planck Institute for Psycholinguistics
Multilingual (adjective), 43–44, 84, 185, 187–188, 190–192, 194, 196
  data/information, 40, 44, 81, 192, 195
  language acquisition, 177, 185–186, 190, 194–196
  population, xvii, 185–186, 191–193 (*see also* Multilingual (person): community/group/population; Population)
  resource, 57, 81, 101
  See also Multilingual (person); Multilingualism/Multilinguality
Multilingual (person), 185–195
  child, xii, 190, 194
  community/group/population, xvii, 185–186, 189, 191–193 (*see also* Multilingual (adjective): population)
  participant, 186–187, 191
  speaker, 185–189, 192–193 (*see also* Speaker, types of: bilingual/multilingual)

Multilingual (person) (cont.)
  See also Speaker, types of: bilingual/multilingual
Multilingualism/Multilinguality, xii, xvii, 61, 158, 185–189, 191–192, 196. *See also* Multilingual (adjective); Multilingual (person)
  assessment of, 156–158, 186, 192, 195
  questionnaire, 156, 189, 192, 195
Multilingual Linked Open Data for Enterprises (MLODE), x, 12, 39–40

Named entity, 25, 59, 100
Namespace, 49, 112–113, 120, 124
Natural Language Processing (NLP), x, xv, 10, 12, 14, 26–27, 29, 34, 36, 42–44, 47, 50–52, 59, 62, 69, 134, 173
Neo-Aramaic, 44
NLP. *See* Natural Language Processing

OAI-PMH. *See* Open Archive Initiative's Protocol for Metadata Harvesting
OBP. *See* ISO; Online Browsing Platform
ODIN. *See* Online Database of Interlinear Text
OKFN. *See* Open Knowledge Foundation
OLAC. *See* Infrastructure, types of: OLAC; Open Language Archives Community
OLiA. *See* Ontologies of Linguistic Annotation
Online Browsing Platform. *See* ISO; Online Browsing Platform
Online Database of Interlinear Text (ODIN), 20–21
OntoLex. *See* Ontology-Lexica Community Group
OntoLingAnnot, xv, 153, 173. *See also* GOLD; OLiA; Ontology: linguistic; OntoTag
Ontologies of Linguistic Annotation (OLiA), xv, 34, 53. *See also* GOLD; OntoLingAnnot; Ontology: linguistic; OntoTag
Ontology, xi, xiii, xv–xvi, 6, 10, 12–14, 20–21, 39, 53, 57, 62, 70, 77, 88, 91, 110, 113, 133, 152–153, 173–174, 176
  linguistic, xv–xvi, 20, 57, 153, 173 (*see also* GOLD; OLiA; OntoLingAnnot; lemon; OntoTag)
  relationship, 110, 113
  vocabulary, 13–14 (*see also* Vocabulary; Vocabulary, types of)
  See also Annotation, types of: ontological; Label, types of: ontological; Ontology-Lexica Community Group; Resource, types of: ontological
Ontology-Lexica Community Group (OntoLex), 12, 39. *See also* lemon; Lexical resource; Lexicon; Ontology
OntoTag, xv. *See also* GOLD; OLiA; OntoLingAnnot; Ontology: linguistic
Open, ix–x, xii, xv, 1, 3, 7, 10–11, 13, 60–61, 75, 89, 131, 134, 201–203, 206–207, 209
  data (*see* Open Data)
  format, 10, 211
  Knowledge Foundation (*see* Open Knowledge Foundation)
  Knowledge International (*see* Open Knowledge International)
  license, 6–7, 11, 50, 59–61
  resource, 1, 11, 60
  science, xvii, 3, 201, 206, 209 (*see also* E-science)
  source, 1–2, 7, 132
  See also Openness
Open Archive Initiative's Protocol for Metadata Harvesting (OAI-PMH), 105, 118–119, 121, 124–125, 133
Open Data, ix–x, xii, xv–xvi, 1–3, 6–7, 50, 60–61, 132, 153, 173, 195–196, 202, 209. *See also* Linguistic Linked Open Data; Linked Open Data)
Open Knowledge Foundation (OKFN), 11, 59, 127
  working group, the, 11 (*see also* Open Linguistics Working Group)
  See also Open Knowledge International
Open Knowledge International, x, 11. *See also* Open Knowledge Foundation
Open Language Archives Community (OLAC), xiv, xvii, 101, 117–122, 124–128, 132, 152–153
  infrastructure, 107, 117, 121, 127
  See also Format, types of: OLAC; Metadata standard: OLAC; Vocabulary, types of: OLAC-specific

Open Linguistics Working Group (OWLG), x, 11, 39–40, 50, 52, 59–61, 63, 153
Openness, xv, 1–2, 7, 60, 211. *See also* Open
Open Researcher and Contributor ID (ORCID), 108–109, 127, 207
Open Science Framework (OSF), 192
ORCID. *See* Open Researcher and Contributor ID
OWL. *See* Web Ontology Language; OWL-DL
OWL-DL, 20, 22, 50. *See also* Description Logics; Web Ontology Language
OWLG. *See* Open Linguistics Working Group

PanLex. *See* Panlingual Lexical Translation
Panlingual Lexical Translation (PanLex), 53, 62
PAROLE, 28, 30
Parse, 32, 142
  constituency, 31–32
  dependency, 29, 32, 133
Parser, 28–29, 133, 147, 172–173. *See also* Grammar
Participant, xiv, 77, 138, 154, 156, 159, 163, 186–188, 190–193, 195, 206
Part of speech (POS), 25, 27–29, 34–35, 44–45, 53, 70, 75–76, 79–83, 89–90, 93
  annotation, 27–29 (*see also* Annotation, types of: morphosyntactic; Tag: part-of-speech)
  tag (*see* Tag: part-of-speech)
  tagger, 29, 132–133, 173 (*see also* Tagger)
PDF. *See* Portable Document Format
Penn Treebank, xv, 27, 29, 35, 44–45. *See also under* Format, types of
People, x, 4, 7, 40, 42, 47, 49, 61, 107, 121, 126, 127, 185–187, 196
  people's name (*see* Personal name)
  See Individual (person); Multilingual (person); Participant; Person (human); Speaker, types of; Subject (participant in study)
Permanent identifier (PID), 132–133. *See also* Persistent identifier
Persistent identifier (PID), 3, 10, 93–94, 104, 107–110, 211. *See also* Permanent identifier; Uniform Resource Identifier
Person (grammatical), 168, 171, 190–191
Person (human), 35, 79, 107, 109–111, 126–127, 135, 148, 151, 185–186, 188. *See also* Individual (person); Multilingual (person); Participant; People; Speaker, types of; Subject (participant in study)
Personal name, 86, 108, 126. *See also* Individual (person); Metadata, types of: person-related; Person (human)
PHOIBLE, 50, 53, 57–58
Phoneme, 53, 57–59
PID. *See* Permanent identifier; Persistent identifier
Polish, 77, 133
Population, xvii, 10, 42, 138, 146, 172, 177, 185–188, 191–192, 195. *See also under* Multilingual (adjective)
Portable Document Format (PDF), 3. *See also* Format; Format, types of
Portuguese, 133
POS. *See* Part of speech
POST, 133, 137–139. *See also* CLAN; MEGRASP; MOR; Part of speech; POST; Tagger
Practice, xvii, 2, 19, 31–34, 73, 77, 96, 126–128, 136, 155, 177, 202–203, 205. *See also* Best Practice; Clinical: Practice; Community: of practice
Preservation, 7, 10–11, 51, 87, 111, 166, 209
  of data, xiv, 101, 154, 202 (*see also* Data management, processes for: preservation)
Privacy, 2, 209. *See also* Data, types of: private; Private data
Private data, xiii, 3, 42, 104. *See also* Data, types of: closed; Data, types of: private; Privacy
Profile, 77, 85–86, 102, 104, 106–108, 110–111, 120, 135, 145, 148, 187, 191–192
  CMDI (*see* Component Metadata Infrastructure: profile)
  language/linguistic, 132, 186, 188–189, 191–192
  See also under Language/linguistic resource; Metadata
Psycholinguistic, xi, 74, 101, 131

Quad. *See* Resource Description Framework: quad; Resource Description Framework: statement; Resource Description Framework: triple
QuantHistLing. *See* Quantitative Historical Linguistics

Quantitative Historical Linguistics (QuantHistLing), 53, 56, 58, 61
Quechua, 193
Query, 158, 161, 163–164, 167–169, 171

RA. *See* Registration Authority
RDF. *See* Resource Description Framework
Recommendation, xii, xvi, 10, 19, 21, 112, 120. *See also* Best practice; OLAC; TEI; W3C
Redundancy, 6, 8, 73, 77, 79–80, 83, 86, 88, 94. *See also* Duplication
Registration Authority (RA), 69, 74, 76, 95. *See also under* International Organization for Standardization
Registry, xv–xvi, 34, 69, 73–74, 78–79, 89, 92–94, 104–105
  Component (*see* Common Language Resources and Technology Infrastructure: Component Registry)
  Concept (*see* Common Language Resources and Technology Infrastructure: Concept Registry)
  Data Category (*see* Data Category Registry; Data Category Repository)
  ISOcat (*see* ISOcat)
  relation, 99, 110, 113 (*see also* RELcat)
Relative clause, 158–164, 177
RELcat, 110–113. *See also* Registry: relation
Replicability, xv, vxii, 1–3, 33, 61, 154, 156, 160, 167, 172, 175, 177, 192, 195
Repository, xvii, 51, 62, 69, 78, 92, 118, 124–125, 131, 140, 203, 206–207, 209–210
  data category (*see* Data Category Repository; DatCatInfo)
  HTTP-accessible, 51
  metadata (*see* Metadata: repository)
  terminology (*see* Terminology: repository)
Reproducibility, 1–3
Research, xi, 100, 155–156, 158, 162, 185–186
  data, xvi–xvii, 99–100, 102, 108–110, 113, 132, 171, 173, 175, 201–202, 209, 211
  hypothesis, 152, 156, 159–161, 168–169, 171–172, 211
  infrastructure, xiv, xvi, 99–101 (*see also* CLARIN; Infrastructure)
  library, xvii, 201–202, 209 (*see also* Library)
  question, ix, xii, 100, 152, 158, 161, 165, 168, 172
  tool, xvi, 99–100, 153, 155, 172, 185, 195–196 (*see also* Tool; Tool, types of)
  See also Reproducibility
Research data, xvi–xvii, 2, 99–100, 102, 109–110, 113, 132, 171, 173, 175, 201–202, 209, 211. *See also* arXiv
Researcher, ix–xii, xiv, 3, 14, 52–53, 56, 59, 61, 70–71, 73, 76, 99, 101, 108, 110, 113, 127, 132–133, 138, 140, 152–156, 159, 161–165, 167–168, 173–174, 176–177, 186–187, 189–193, 195–196, 205, 207
ResearcherID, 108–109
Research, types of, 40, 185
  child language (*see* Child language)
  collaborative, xi, xvii, 151, 154–155, 177
  community, 185 (*see also* Community)
  cross-linguistic (*see* Cross-linguistic)
  data-intensive, 40
Resolvability, 4, 48, 51, 109, 112. *See also* Accessibility; Addressability; Findability; Uniform Resource Identifier: resolvable
Resource, ix–x, xiii–xiv, xvi–xviii, 2, 4, 7–13, 19–20, 25–26, 28, 34, 40, 42–44, 47–53, 57, 58–62, 69, 72, 76–78, 82, 99–102, 106, 109–110, 113, 117–123, 125–128, 131, 133–134, 142, 156, 158, 168–169, 172–173, 202, 210
  accessibility, 42, 49, 100 (*see also* Accessibility)
  creator, 101, 107, 110
  description, 101–102, 112, 118, 126 (*see also* Language/linguistic resource: description [of]; Resource Description Framework)
  development, 40, 47, 52
  integration, xvi, 6–7, 39–40, 47, 62
  metadata, 57, 102, 114, 126 (*see also* Metadata)
Resource Description Framework (RDF), xii, xvii, 4–11, 14, 48–49, 51, 53, 59, 61–62, 100, 107, 111, 114, 121–128
  -based language, 14, 35, 51
  -based representation, 5, 111, 114
  data, 4, 49 (*see also* Resource Description Framework: resource)
  format, 10, 14, 35

graph, 53, 57, 121
model, 53, 56–57
object, 5, 48–49, 121–122, 126
predicate, 48–49, 121–122
property, 5, 48–49, 121–124, 126, 128
quad, 7 (*see also* Resource Description Framework: statement; Resource Description Framework: triple)
representation, 111–112, 114
resource, 4, 5, 48–49, 124–126 (*see also* Resource Description Framework: data)
schema, 5–6, 124 (*see also* Schema)
serialization, 4–5, 14
statement, 5, 7, 48, 121, 125–126 (*see also* Resource Description Framework: quad; Resource Description Framework: triple)
subject, 5, 48–49, 121–122
technology, 7, 9, 14, 62, 111
triple, 5, 7, 48–49, 111, 121–122, 173 (*see also* Resource Description Framework: quad; Resource Description Framework: statement)
Resource, types of, ix, xi, 4, 9, 11–12, 19, 27, 33, 41, 47, 51, 57, 66, 70, 81, 88, 90, 92, 94, 101, 120–121, 148, 173–174, 188
CLLD (*see* Cross-linguistic Linked Data)
community-supported, 205 (*see also* Community)
digital/electronic/electronically encoded, 19, 42–43, 73, 99, 101
language (*see* Language/linguistic resource; Resource, types of: language-related)
language-related, 100, 102, 112, 114 (*see also* Language/linguistic resource)
lexical (*see* Lexical resource)
library (*see* Library: resource; Resource: description)
Linguistic Linked Open Data, ix, 12–13, 15, 40, 51, 172, 174 (*see also* Linguistic Linked Open Data)
LLOD (*see* Resource, types of: Linguistic Linked Open Data)
LOD (*see* Resource, types of: Linguistic Open Data)
online, xvi–xvii, 69, 209
ontological, 14 (*see also* Ontology)
open (*see* Open: resource)
open language (*see under* Language/linguistic resource)
RDF (*see* Resource Description Framework)
sharable, 173 (*see also* Share/sharing)
TalkBank (*see* TalkBank)
terminology (*see* Terminology: resource)
TermWeb (*see* TermWeb)
URI (*see* Uniform Resource Identifier)
web (*see* Web: resource)
Responsibility, 1, 21, 95, 112, 202, 208
Reusability, x, xv–xvi, 1–3, 7, 9, 10, 25, 30–31, 33, 40, 47, 49–50, 61, 102, 125, 196, 201, 205, 209, 211
of data (*see* Data management, processes for: reuse [of])
of language/linguistic resources (*see* Language/linguistic resource: reuse [of])
See also Accessibility; FAIR principles, Findability; Interoperability; Share/sharing
Right, 2, 11, 118, 156. *See also* Copyright; License
Russian, 41–42

Schema, 102, 104, 106. *See also* Schema.org; Scheme
XML (*see* XML Schema)
RDF (*see* Resource Description Framework)
Schema.org, 110, 112
Scheme, 20, 102
annotation (*see* Annotation: scheme)
metadata (*see* Metadata)
See also Schema
Scholarship, ix–xi, 151, 201–202
Science, ix, xii, xiv, 1–2, 39, 95, 151, 173, 201, 206, 211
Science, types of, xiii–xiv, 61
cognitive, ix, 61, 151–152, 177
computer, x, xv, 203
information, xi, 173, 206
language, xi, xiv, 10, 62, 151, 155, 172, 176–177
open, xvii, 3, 201, 206, 209
social, ix, xii, xvi–xvii, 26–27, 99–100, 151, 156
Scientist, xii, 40, 152, 201–202, 206–207, 209–211. *See also* Researcher

SDSS. *See* Sloan Digital Sky Survey
Segmentation, 3, 31
Semantic role, 25, 33–35, 45. *See also* Annotation, types of: semantic role; Label, types of: semantic role
Semantic Web, x, xii, xvi, 11, 19–20, 35, 40, 49, 51, 53, 56, 59, 61, 99–100, 105, 107, 110–112, 120–121, 172–174, 176
  dataset, xvii, 11, 56, 100 (*see also* Dataset)
  technologies, 40, 51, 56, 61, 107
  See also Linked Data; Linked Open Data
Sentence, 167, 169, 171, 177
Session, 137, 156, 161, 166, 169, 171, 174–175
SGML. *See* Standard Generalized Markup Language
Share/sharing, 20, 40, 50–52, 59, 100, 102, 109. *See also* Data management, processes for: sharing [of]; Data, types of: sharable/shared; Infrastructure, types of: shared; Language/linguistic resource: sharing [of]; Resource, types of: sharable; Reusability; Vocabulary, types of: shared
Simple Knowledge Organization System (SKOS), 106, 109–110, 113, 122–123
SketchEngine, 134. *See also* Corpus: tool; Tool, types of: corpus; Tool, types of: cybertool
SKOS. *See* Simple Knowledge Organization System
Sloan Digital Sky Survey (SDSS), 201
SMC Browser, 104, 106, 112. *See also* Tool, types of: web-based
Spanish, 28, 41, 76, 133, 142, 158, 165–169, 171–172, 174, 190
SPARQL, 4, 6–8, 14, 49–50, 57, 107, 111, 114, 128
  endpoint, 7–8, 49–50, 57, 59, 111, 114, 128
Speaker, types of, 41–43, 50, 131, 135, 165, 169, 185–192
  bilingual/multilingual, 185–189, 190–193 (*see also* Multilingual (adjective); Multilingual (person))
  monolingual, 186, 190–191, 186, 190–191, 195
Speech-Language Pathologist (SLP), 136, 140, 142, 145, 147–148
SQL. *See* Structured Query Language
Standard, xii–xiv, xvi–xvii, 2–4, 6–7, 9–10, 19–20, 27, 30–32, 34, 44, 48–49, 51, 69–78, 82, 90–92, 94, 111, 117–118, 121–122, 126–128, 161, 173–174
  data category (*see* Standard: ISO 12620:2009; Data category)
  encoding, 101, 107
  format, 57
  framework, 113
  international, 100, 108, 112
  ISO, 112 (*see also* International Organization for Standardization)
  ISO 639–3:2007, 44, 109, 120, 122, 126 (*see also* Language/Linguistic: metadata; Code)
  ISO 3166:2013, 99–100, 109
  ISO 8601:2013, 109
  ISO 12620:2009, 102, 106 (*see also* Standard: data category; ISOcat)
  ISO 15836–1:2017, 101
  ISO 24622–1:2015, 108, 111
  ISO 27729:2012, 108
  language, 44, 48–49, 59 (*see also* eXtensible Markup Language; IGT; OWL; RDF; SPARQL)
  W3C, 48, 59 (*see also* World Wide Web Consortium)
  web protocol, 51
  XML, 104 (*see also* eXtensible Markup Language)
  See also CES; Data management: standard; Harmonization; International Organization for Standardization; International Standard Name Identifier; Library: standard; Metadata standard; Standard Generalized Markup Language
Standard Generalized Markup Language (SGML), 30, 71–72
Standardization, xii, xiv, xvi, 4, 9, 34, 42–44, 49, 62, 69–71, 73–77, 82, 90, 92, 94, 100, 107, 118, 127, 135, 138–139, 145, 153–154, 161–162, 173, 176, 189, 196. *See also* Harmonization
Standoff, 26, 31, 53
Structured Query Language (SQL), 9, 49
Student, xiv, x, 61, 154–156, 158, 187, 211
Subject (grammatical function), 160, 162, 165, 167, 190

Subject (node), 5, 48–49, 121–122, 125–126
Subject (participant in study), xiii, 156, 159, 161, 165–169, 171–172, 174–175, 188, 193
Subject (RDF triple). *See* Resource Description Framework: subject
Subject (topic), 39, 79, 89, 118–120, 126, 206–208
Sustainability, xiv, 8, 25, 101, 109, 128, 154, 201–203, 205, 207–211. *See also* Infrastructure, types of: sustainable; Maintainability
Swedish, 28
Syntactic, xii, 99, 134, 159–160, 188, 193
  analysis (*see* Analysis, types of: morphosyntactic)
  annotation (*see* Annotation, types of: syntactic)
  constituency, 31 (*see also* Parse: constituency)
  elements, 140, 164 (*see also* Syntactic: units)
  information (*see* Information; Information, types of)
  format, 34 (*see also* Format, types of: CoNLL; Format, types of: Penn Treebank)
  function, 29, 165, 167
  interoperability (*see* Interoperability, types of: syntactic; Interoperability)
  parser (*see* Parser)
  realization, 34
  treebank, 28
  units, 158 (*see also* Syntactic: elements)

Tag, 20–21, 25, 28–31, 35, 44, 59–60, 87–88, 133, 165, 196
  part-of-speech, 25, 28–29, 35, 44, 133, 196 (*see also* Annotation, types of: part-of-speech)
  sense, 25 (*see also* Annotation, types of: sense)
Tagalog, 190
Tagger, 29, 100, 132–133, 137–138, 173
  morphosyntactic (*see* Part of speech: tagger)
  part-of-speech (*see* Part of speech: tagger)
TalkBank, xvii, 131–134, 142, 147. *See also* Corpus
TBX. *See* TermBase eXchange
Technology, x–xvi, 2–3, 7, 9, 13–14, 22, 39–40, 47, 61, 100, 104, 107, 111, 133, 153–155, 174, 177, 178, 202, 205, 209, 211
TEI. *See* Text Encoding Initiative
Term bank, 71–72. *See also* Termbase; Terminology: repository
Termbase, 7, 13, 69–73, 77–78. *See also* Term bank; Terminology: repository
TermBase eXchange (TBX), 70, 72, 78, 84–87, 92
Terminology, xii, xvi, 6, 10, 13, 19, 34, 49, 51, 61, 69, 71–73, 77–78, 80–82, 90, 92–93, 95, 112, 173
  database, 69 (*see also* Termbase)
  interchange, 10 (*see also* MAchine-Readable Terminology Interchange Format)
  management [system], xvi, 71, 73, 78 (*see also* TermWeb)
  repository, 49, 51, 61 (*see also* Term bank; Termbase)
  resource, 73, 80, 92 (*see also* Resource; Resource, types of)
  service, 112 (*see also* CLAVAS)
TermWeb, xvi, 77–78, 84, 85–87, 89, 92–94. *See also* Terminology: management [system]
Text Encoding Initiative (TEI), 19–20, 30, 39, 88, 101–102
Thai, 133
Theory, x, xiii, 1, 25, 27–29, 39, 43, 56–57, 134, 152, 164–165, 185, 211
Therapy, 135, 137, 147
Thesaurus, 19, 60
Token, 29, 31–32
Tokenization, 31–33
Tool, xvi, xvii, 6, 10, 14, 20, 25, 31–32, 35, 40, 42, 46–47, 50, 52, 57, 61–62, 78–79, 87, 99–100, 106
Toolbox, 44, 47
Tool, types of, xiv, 6, 14, 25, 43, 78, 87, 99–100, 104, 111–112, 139, 151–152, 154, 192–193, 202
  analysis, 134, 151, 172, 176
  assessment, 192, 195, 207
  clinical, 139 (*see also* KIDEVAL)
  CMDI metadata editor (*see* COMEDI; Component Metadata Infrastructure)
  corpus, 134 (*see also* SketchEngine)
  cybertool, xiv, 151, 154, 160, 176–177 (*see also* Data Transcription and Analysis tool; EVAL; KIDEVAL; SketchEngine)
  NLP, 34, 42, 46, 52 (*see also* Natural Language Processing)

Tool, types of (cont.)
  named entity recognizer (*see* Annotation, types of: named entity; Named entity)
  research (*see* Research)
  tagger (*see* Tagger)
  web-based, 100, 104 (*see also* SMC Browser)
Transcript, 44, 131, 134, 136–137, 147–148, 167, 174–176, 193, 196
  file, 132–133, 137
Transcription, 19, 25, 44, 57, 109, 132, 136–137, 146–147, 153, 155–156, 159, 163–166, 168–169, 174, 176, 194–196
  DTA tool's (*see* Data Transcription and Analysis Tool)
Translation, 26, 43, 44, 46, 50, 52–53, 56, 58, 61–62, 69, 78, 134, 189
  graph, 44, 56–57, 62 (*see also* Graph)
Triple, 57. *See also* Resource Description Framework: quad; Resource Description Framework: statement; Resource Description Framework: triple
Tulu, 162
Turkish, 28
Turtle, 4–5, 49. *See also* Resource Description Framework

Ugaritic, 52
Uniform Resource Identifier (URI), 4–5, 16, 48–49, 51, 57, 107–109, 111–112, 114, 121–122, 124–127
  persistent, 109, 114 (*see also* Persistent identifier)
  resolvable, 51 (*see also* Resolvability)
  resource, 107–108, 121
Uniform Resource Locator (URL), 4–5, 10, 59, 87, 92
Urdu, 190
URI. *See* Uniform Resource Identifier
URL. *See* Uniform Resource Locator
Usability, 26, 78, 106, 152, 176, 201–202, 209, 211
Utterance, 26–27, 131, 136, 140, 153, 159, 161–169, 171–172, 174, 176, 192–194, 196
  Mean length of (*see* Mean length of utterance)

VCLA. *See* Virtual Center for Language Acquisition Research
VIAF. *See* Virtual International Authority File (VIAF)
Virtual Center for Language Acquisition Research (VCLA), xv, 155, 161, 172, 177–178
Virtual International Authority File (VIAF), 108. *See also* Authority file
Virtual Language Observatory (VLO), 99, 105–106, 110–111, 113–114, 133, 135
Virtual Linguistic Lab (VLL), 151–152, 155–156, 158
VLL. *See* Virtual Linguistic Lab
VLO. *See* Virtual Language Observatory
Vocabulary, xv, 6, 9, 49–50, 99, 106, 109–113, 122–123, 125, 133, 135, 143, 189–190, 211
  documentation, 123 (*see also* Documentation)
  CLARIN (*see* Common Language Resources and Technology Infrastructure: vocabulary)
  CLAVAS (*see* CLAVAS)
  CMDI (*see* Component Metadata Infrastructure: vocabulary)
  CMDI-metadata (*see* Component Metadata Infrastructure: metadata)
  controlled, 99, 102, 109
  Linked-Data (*see* Linked Data)
  OLAC Field, 126 (*see also* Vocabulary: OLAC-specific)
  OLAC Role, 126 (*see also* Vocabulary: OLAC-specific)
  OLAC-specific, 126 (*see also* Vocabulary: OLAC Field; Vocabulary: OLAC Role; Vocabulary: OLAC Type; Open Language Archives Community)
  OLAC Type, 126 (*see also* Vocabulary: OLAC-specific)
  RDF (*see* Resource Description Framework)
  service, 109, 113 (*see also* Vocabulary: CLAVAS)
  shared, xv, 50, 52–53, 109, 117–118, 120–122 (*see also* Share/sharing)
  XML, 88 (*see also* eXtensible Markup Language)
Vocabulary, types of, xii, 6, 9–10, 13–14, 49, 58, 109–113, 122–123 (*see also* Infrastructure; Metadata; Ontology)

W3C. *See* World Wide Web Consortium
WALS. *See* World Atlas of Language Structures
Web, xvii, 4, 20, 48, 51, 56, 72, 105, 110, 133, 147, 153, 155–156, 196
  resource, 14, 48, 50–52, 118 (*see also* Resource)
  service, 14, 32, 34, 105, 109, 112–113 (*see also* CLAVAS; CMDI2DC)
  standard (*see* Standard: web protocol)
Web Ontology Language (OWL), xii, 6, 20–22, 35, 50–51, 61–62, 110
  format (*see* Format, types of)
  See also Ontology; OWL-DL
Web, type of, 4–5
  of Data, 4–6, 7, 9, 11, 14, 39, 48, 51, 117, 120–121, 128, 174
  of Documents, 4–5
  Semantic (*see* Semantic Web)
  World Wide Web (WWW), xii, 20, 72, 120
Word, 3, 19, 21, 27–28, 33, 39, 41, 43–44, 46, 50, 53, 56, 59, 60, 62, 70, 101, 123, 137–138, 153–154, 166–167, 192–194, 196. *See also* Form
Wordlist, 19, 21, 39, 43–44, 46, 53, 56, 62
WordNet, 59, 123, 134
World Atlas of Language Structures (WALS), 41, 58
World Wide Web Consortium (W3C), xii, 4, 9–10, 12–13, 20, 39, 48, 120, 127, 174
WWW. *See* Web, type of: World Wide Web

XML. *See* eXtensible Markup Language
XML Schema, 30, 118, 124
  Definition (XSD), 104
XSD. *See* XML Schema: Definition

Yiddish, 190–191, 193